FACULTAD DE CIENCIAS

GRADO EN BIOLOGÍA

TRABAJO FIN DE GRADO

CURSO ACADÉMICO [2020-2021]

TÍTULO:

MINERÍA DE DATOS APLICADOS A DATOS BIOLÓGICOS

AUTOR:

JOSÉ ARMANDO VIDAL MIRALLES

UNIVERSITAT D'ALACANT
Facultat de Ciències
Facultad de Ciencias

FACULTAT DE CIÈNCIES

GRAU EN BIOLOGIA

TREBALL DE FI DE GRAU

CURS ACADÉMIC [2020-2021]

TÍTOL:

MINERIA DE DADES APLICADA A DADES BIOLÒGIQUES

AUTOR/A:

JOSÉ ARMANDO VIDAL MIRALLES

UNIVERSITAT D'ALACANT
Facultat de Ciències
Facultad de Ciencias

FACULTY OF SCIENCES

DEGREE IN BIOLOGY

FINAL PROJECT

ACADEMIC YEAR [2020-2021]

TITLE:

DATA MINING APPLIED TO BIOLOGICAL DATA

AUTHOR:

JOSÉ ARMANDO VIDAL MIRALLES

Resumen

En diciembre de 2019 se detectó una nueva enfermedad respiratoria aguda, la enfermedad denominada como COVID-19 (*coronavirus disease*) en la ciudad de Wuhan, provincia de Hubei, China. Desde el momento en el que se decretó la enfermedad como pandémica en marzo de 2019, los gobiernos con competencias en la gestión de la pandemia han impuesto diferentes medidas para mitigar la propagación del virus SARS-CoV-2.

Esta situación demanda la necesidad de conocer con la máxima precisión posible la evolución de la enfermedad en cada región para así tomar las mejores decisiones en la gestión de la pandemia.

Una de las metodologías más usadas durante la pandemia para predecir la propagación de la enfermedad fue la construcción de redes neuronales artificiales (ANN). Con este método, basado en el aprendizaje profundo, se ha realizado una predicción de la COVID-19 de diferentes municipios de la Comunidad de Madrid.

Las predicciones a partir de redes neuronales podrían cambiar drásticamente la gestión de esta y futuras pandemias, dejando entrever la posibilidad de que las medidas sean más específicas para cada región.

Abstract

In December 2019, a new acute respiratory illness, the disease named as COVID-19 (coronavirus disease) was detected in Wuhan city, Hubei province, China. From the time the disease was decreed as pandemic in March 2019, governments with powers in pandemic management have imposed different measures to mitigate the spread of SARS-CoV-2 virus.

This situation demands the need to know as precisely as possible the evolution of the disease in each region in order to make the best decisions in the management of the pandemic.

One of the most widely used methodologies during the pandemic to predict the spread of the disease was the construction of artificial neural networks (ANN). With this method, based on deep learning, a prediction of the COVID-19 of different municipalities of the Community of Madrid has been made.

Predictions from neural networks could dramatically change the management of this and future pandemics, raising the possibility of more region-specific measures.

Índice

1.Introducción

1.1. Precedentes de otros coronavirus

Los coronavirus son un grupo de virus que pueden llegar a infectar a un gran grupo de animales en el cual se encuentran los seres humanos y causan desde infecciones respiratorias leves a graves, pudiendo ocasionar la muerte. En los años 2002 y 2012, surgieron dos tipos de coronavirus de origen zoonótico: el coronavirus del síndrome respiratorio agudo grave (SARS-CoV) y el coronavirus del síndrome respiratorio de Oriente Medio (MERS-CoV), respectivamente. Estos coronavirus provocaron enfermedades fatales que convirtió a los nuevos coronavirus emergentes en un nuevo problema para la salud pública mundial en este siglo. (Hu *et al*., 2020). El SARS-CoV-2 ha superado con creces la propagación de los dos anteriores tanto en número de contagios como en área espacial, donde el MERS-CoV apenas alcanzó los 707 casos y 252 muertes en 2014, mientras el SARS-CoV unos 8400 casos detectados y 800 muertes. (Sharif-Yakan y Kanj, 2014).

Estos virus no son nuevos para la comunidad científica, ya que desde la década de 1960 se han identificado nuevos coronavirus que han infectado a la población, causando síntomas leves similares a los resfriados comunes. De las siete especies conocidas de coronavirus, cuatro afectan al tracto respiratorio superior y las otras tres al tracto respiratorio inferior. Estas tres especies son el SARS-CoV, el MERS-CoV y el SARS-CoV-2 y pueden llegar a causar enfermedades respiratorias graves (Uddin *et al*.,2020).

1.2. Descripción del SARS-COV-2

El SARS-CoV-2 es un β-coronavirus muy similar a los dos anteriores nuevos coronavirus descubiertos en el siglo XXI, al SARS-CoV y al MERS-CoV, compartiendo el 79% del genoma con el primero y un 50% con el segundo.

Como muchos otros coronavirus, el SARS-CoV-2 contiene ARN monocatenario con sentido positivo y envuelto con un genoma de alrededor de 30 kb. El genoma viral codifica 16 proteínas no estructurales que son necesarias para la replicación y patogénesis del coronavirus, cuatro estructurales entre las que se encuentran la envuelta, la nucleocápside, la membrana y las glicoproteínas. Esta última es fundamental para la tipificación del virus, la respuesta a nuevas vacunas y otros nueve factores (Uddin *et al*., 2020). Su genoma contiene seis marcos de lecura abierta (ORF) principales propios de otros coronavirus. (Wieczorek, Siłka y Woźniak, 2020)

1.3. COVID-19

La enfermedad COVID-19 (*coronavirus disease 2*) fue descrita en diciembre de 2019 en Wuhan, China, y es causada por el virus descrito en enero del 2020 SARS-CoV-2(*severe acute respiratory síndrome coronavirus 2*) (Ching-Cheng *et al*., 2020). Esta enfermedad provoca desde infecciones asintomáticas hasta síntomas leves respiratorios, llegando en los casos graves a causar neumonía grave, dificultad aguda para respirar y la muerte (Uddin *et al*.,2020). Entre los principales síntomas se encuentran fiebre, tos, náuseas, dificultad para respirar, fatiga, dolor muscular, dolor de cabeza, perdida del gusto y el olfato, dolor de garganta, secreción nasal y diarrea (Burke *et al*.,2020)

Se piensa que la principal forma de transmitir la enfermedad es a través de gotitas respiratorias, lo que provoca que la mayoría de contagios se produzcan en espacios cerrados o con poca ventilación. El tiempo de incubación dura normalmente entre tres y siete días, pero puede durar incluso dos semanas. (Wieczorek, Siłka y Woźniak, 2020)

La rápida propagación de la COVID-19 y las miles de muertes causadas entonces llevaron a la Organización Mundial de la Salud a declarar la enfermedad como pandémica el 12 de marzo de 2020 (Ciotti *et al*., 2020).

Hasta la fecha, se han registrado alrededor de 185 millones de casos y más de 4 millones de muertes en todo el mundo, siendo en España más de 80,000 muertes y casi 4 millones de casos confirmados. (OMS,2021)

1.4. Situación en España

En el caso de España, la gestión ha sido independiente en la mayoría de criterios por parte de las comunidades autónomas y todavía sigue siendo gestionada la evolución de esta forma. Este hecho, además de las diferencias entre factores que pueden cambiar la evolución de la pandemia entre las diferentes CCAA como la edad media de la población, la temperatura o la densidad de la población provoca que existan diferencias importantes per se entre las CCAA (Medeiros *et al.* 2020).

Durante la pandemia, la comunidad con mayor incidencia ha sido la Comunidad de Madrid, teniendo la mayor incidencia acumulada de España con alrededor de un tercio de los casos en mayo de 2020(Condes y Arribas, 2021) y la de mayor número de casos totales con unos 750,000 casos (Datos Comunidad Madrid,2021).

1.5. Situación en la Comunidad de Madrid

La Comunidad de Madrid es la tercera comunidad más poblada de España después de Andalucía y Cataluña, con 6,779,888 habitantes y donde se encuentra la capital de España, Madrid. Demográficamente, la región tiene la densidad de población más elevada del país con 844.53 habitantes/km^2, seguida de Cataluña con 743.19 habitantes/km^2 (INE,2020; Madrid, 2020). Ser la segunda comunidad más importantes del PIB con el 19% y el centro económico del país junto con Cataluña, hacen de la Comunidad de Madrid un interesante objeto de estudio para conocer la evolución de la COVID-19 y determinar factores que puedan ser relevantes para la propagación de la enfermedad (INE, 2020).

Su paisaje urbanístico permite denominar a la ciudad de Madrid, junto con Berlín y Londres, como "ciudades fragmentadas". Esta denominación realizada por la Agencia Europa de Medio Ambiente significa que estas ciudades tienen áreas verdes inconexas, presentando superficies impermeables. Esta estructura urbanística favorece a la población por múltiples motivos tanto para la salud mental como física, aportando áreas al aire libre que reducen espacios potenciales para la transmisión de la enfermedad, así como su propia distribución (Menéndez e Higueras, 2020).

Respecto a sus datos, ha estado por encima de la media del resto de comunidades en incidencia acumulada durante toda la pandemia salvo en el último ascenso de casos a finales de junio de este año, considerando la incidencia acumulada como el número de personas diagnosticadas en los últimos 14 días por cada 100,000 habitantes (Figura 1).

En relación a la evolución que ha tenido, se puede decir que no ha seguido una evolución muy diferente a la del resto de CCAA, aunque existe una mayor amplitud en la función de la comunidad madrileña, como se observa en la Figura 2 (Epdata, 2021).

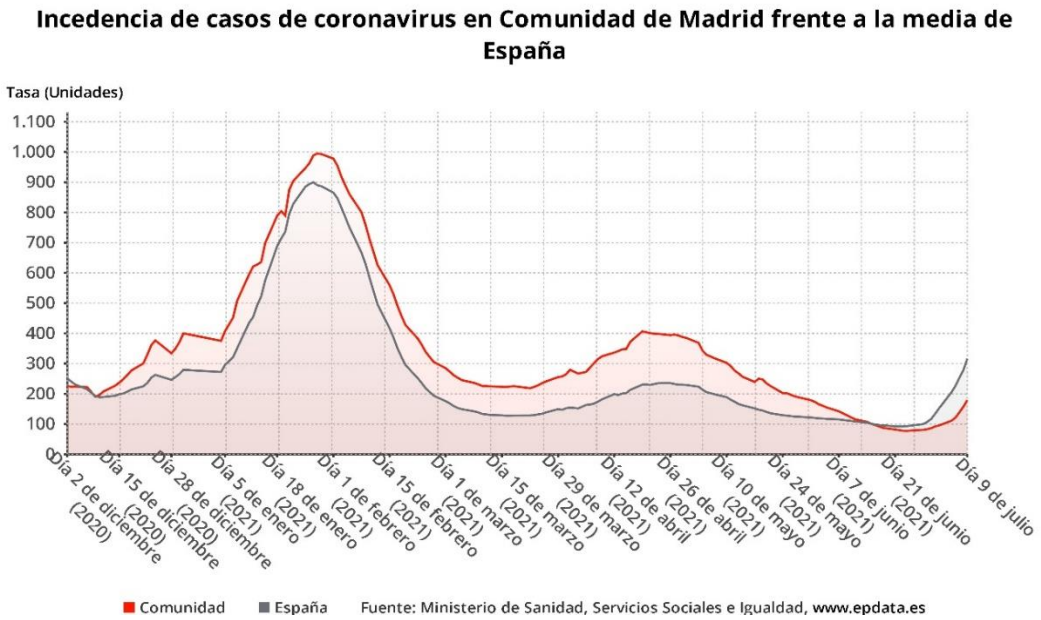

Figura 1.-Incidencia de casos de coronavirus en la Comunidad de Madrid frente a la media del resto de CCAA. La zona roja representaría la incidencia de casos de la Comunidad de Madrid y la gris la de la media del resto de las CCAA (Epdata, 2021).

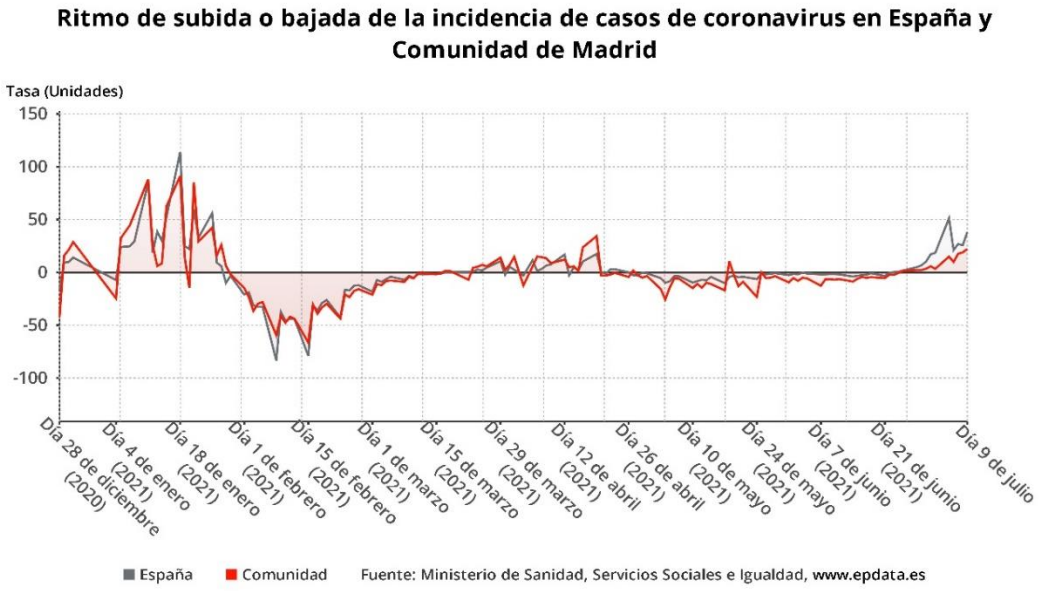

Figura 2-. Aumento o disminución de la incidencia de casos en la Comunidad de Madrid y en el resto de CCAA (Epdata, 2021).

1.6. Predicciones anteriores

Antes de que se recopilasen los suficientes datos como para realizar una predicción a partir de redes neuronales, se utilizaban otros métodos como predictores estocásticos que funcionarían únicamente para una región o país. Utilizando estos métodos y con tan

pocos datos, darle más complejidad al modelo añadiendo datos de lugares diferentes los haría todavía más heterogéneos y no ayudaría al modelo (Ying, 2019; Wieczorek, Siłka y Woźniak, 2020).

1.7. Inteligencia artificial

Por inteligencia artificial se refiere a aquella inteligencia que poseen las máquinas, en contraste con la inteligencia natural de los humanos. De esta forma, una posible definición sencilla y general de inteligencia artificial podría ser el esfuerzo para automatizar mediante máquinas tareas intelectuales normalmente realizadas por humanos (Torres, 2018).

1.7.1 Aprendizaje automático

El aprendizaje automático o automatizado (del inglés *machine learning*) es una rama en desarrollo de la inteligencia artificial y en concreto de los algoritmos computacionales que están diseñados para imitar la inteligencia humana y la forma en la cual aprendemos a partir del entorno que le rodea o de los estímulos que recibe (El Naqa y Murphy, 2015). Está basado en algoritmos que permiten al sistema aprender por su propia experiencia. Por ejemplo, al introducir datos el sistema aprende patrones y responde con salidas, aplicando el aprendizaje previo para los resultados o predicciones. De esta forma, el sistema aprende de forma autónoma sin la necesidad de intervención humana gracias a un algoritmo estadístico que aprende y mejora automáticamente cada vez que obtiene nuevos datos o patrones (Sharma, Sharma y Jindal, 2021)

Se originó en la década de 1950 junto al movimiento de IA y se focaliza fundamentalmente en la predicción y en la optimización. (Bi *et al*., 2019).

Es considerado como el nuevo caballo de batalla en el bigdata y ha tenido éxito en diversos campos, abarcando ámbitos multidisciplinares que van desde la biología computacional o aplicaciones biomédicas hasta la ingeniería espacial o las finanzas; demostrando ser una herramienta fundamental en múltiples áreas aparentemente lejanas (El Naqa y Murphy, 2015).

En el ámbito de la epidemiología está demostrado que puede ser una herramienta que puede revolucionar por completo la disciplina en un futuro no muy lejano y en la actualidad ya es una herramienta fundamental para entender la propagación de enfermedades, como se ha podido demostrar con la pandemia de la COVID-19 (Bi *et al.*, 2019).

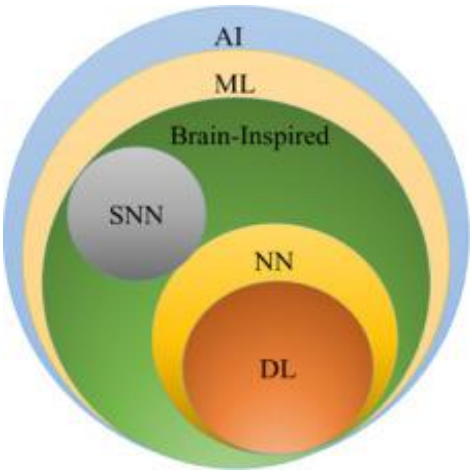

Figura 3-. Correlación entre inteligencia artificial (AI), aprendizaje automático (ML) y aprendizaje profundo (DL). (Sharma, Sharma y Jindal, 2021)

1.7.2. Aprendizaje profundo

El aprendizaje profundo (del inglés *deep learning*) es un subcampo del aprendizaje automático que se basa en el aprendizaje de diferentes niveles de representaciones que corresponden a una jerarquía de características, factores o conceptos donde los conceptos de los niveles superiores están definidos por los de los niveles inferiores, de la misma forma que los de niveles inferiores pueden ayudar a definir los conceptos de niveles superiores (Deng y Yu, 2014). Haciendo uso de una definición más simple, utiliza la composición de una gran cantidad de funciones no lineales para modelar la compleja dependencia entre las características de entrada y las etiquetas (Fan, Ma y Zhong, 2019).

1.7.3. Red Neuronal

Las redes neuronales son consideradas actualmente como uno de los procesos más eficientes a la hora de procesar cantidades enormes de datos que pueden ser analizados para descubrir patrones, tendencias; realizar pronósticos y predicciones, etc. (Tamang, Singh y Datta, 2020)

Una red neuronal se puede definir como un sistema de procesamiento de información no lineal que combinan numerosas unidades de procesamiento con una serie de características como la autoadaptación, la autoorganización y el aprendizaje (Ding *et al.*,2013).

El concepto de red neuronal está basado en las conexiones que se crean entre neuronas en el cerebro humano. Una red neuronal consiste en un gran número de

neuronas que están interconectadas entre sí y separadas por capas. (Mishra y Srivastava, 2014)

La red neuronal se divide en tres tipos de capas principales:

Capa de entrada. Esta capa recibe los datos, que se denominan datos de entrada, para pasarlos a la primera capa oculta.

Capa oculta. Este tipo de capa, en la cual en una red neuronal suele haber más de una para ajustar el modelo a la complejidad necesaria para que la red sea entrenada adecuadamente, realizarán cálculos con los datos de entrada. Uno de los motivos por los cuáles es necesario experimentar en el entrenamiento con diferentes estructuras es la dificultad de encontrar la cantidad de capas y de neuronas ocultas adecuadas para nuestro modelo. Por lo general, en las primeras capas ocultas se introducen el mayor número de neuronas, disminuyendo poco a poco su número hasta llegar a la salida o salidas.

Capa de salida. La última capa devuelve la predicción y existirán tantas neuronas de salida como predicciones de diferentes variables queramos obtener.

Cuantas más capas y neuronas utilicemos, mayor profundidad tendrá nuestro modelo, expresión por la cuál se denomina aprendizaje profundo (Bagnato, 2019).

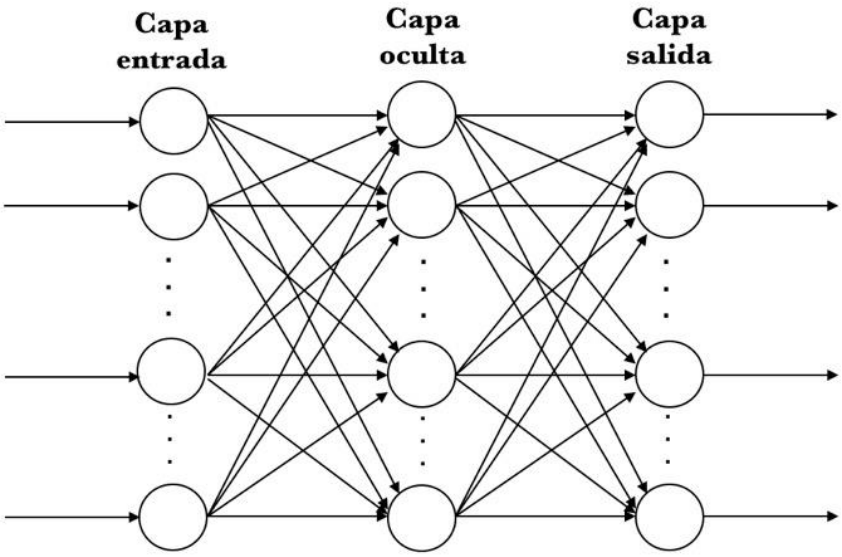

Figura 4-. Esquema de una red neuronal (Torres, 2018)

Cada conexión entre neuronas está asociada a un peso que determinará la importancia de esta conexión al multiplicarse por el valor de entrada. Los pesos iniciales se asignan aleatoriamente por el modelo para después ajustarse de forma autónoma (Bagnato, 2019). La neurona recibe múltiples entradas para finalmente devolver una única salida al resto de neuronas (Krukeja *et al*., 2016).

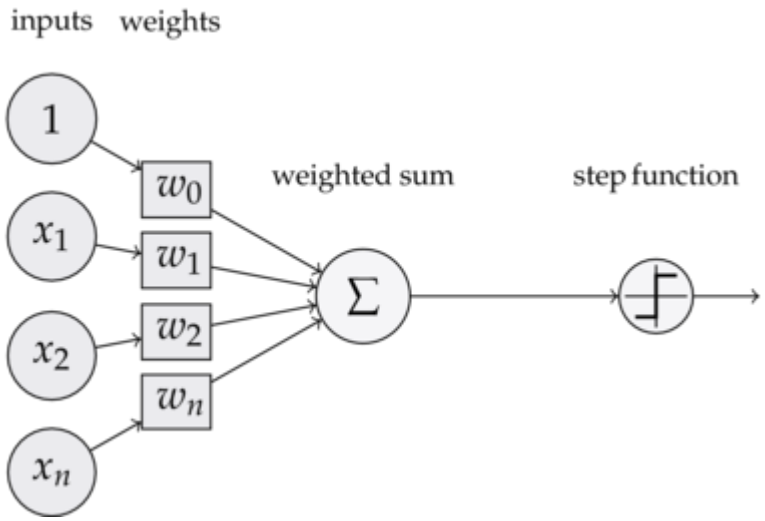

Figura 5-. Esquema de las interacciones entre los pesos en una red neuronal (Estrada, 2016).

Para determinar si una neurona se activa o no, se utiliza la llamada función de activación. Esta función decidirá si la suma de los valores recibidos a una determinada neurona supera un umbral que permite a la neurona activarse y enviar un valor diferente a la siguiente capa (Bagnato, 2019).

2. Objetivos

El principal objetivo de este TFG es realizar una predicción de la evolución de los casos de COVID-19 en diferentes municipios de la Comunidad de Madrid a partir de una red neuronal.

Otro objetivo sería discutir los patrones socioeconómicos que pueden determinar la evolución de la enfermedad. De manera bibliográfica, se intentará buscar estos patrones con casos semejantes y trabajos anteriores en lugares diferentes.

3. Cronograma

El programa a seguir para la confección del trabajo se puede describir detalladamente de la siguiente forma, dividiendo las tareas en los meses posteriores a la elección del tema:

-Marzo. Búsqueda bibliográfica para iniciar la investigación en el estudio del comportamiento de las redes neuronales. Elección del tipo de red neuronal y del conjunto de datos. Determinar factores de interés para estudiar que influyan en la propagación de la COVID-19.

-Abril. Tras la búsqueda bibliográfica, durante el final de marzo y principios de abril se procedió al aprendizaje de la metodología para realizar la red neuronal adecuada para los datos obtenidos. Se procedió a la descarga de los datos.

-Mayo. Una vez se obtuvieron los datos, se procedió a la construcción y entrenamiento de la red basándose en la búsqueda bibliográfica previa. Realización de pruebas para determinar los parámetros ideales para el entrenamiento de la red (número de *epochs* y número de días a predecir). Obtención de resultados.

-Junio. Recolección de bibliografía útil para la discusión de los resultados y las conclusiones, extrayendo de los artículos escogidos factores y ejemplos semejantes al del estudio. Redacción del TFG

-Julio. Continuación en la redacción de la memoria y entrega y posterior defensa.

Para la realización de este TFG se siguió el cronograma representado en la Tabla 1:

Tabla 1-. Cronograma de actividades realizadas durante el TFG.

Tareas	Marzo	Abril	Mayo	Junio	Julio
Búsqueda bibliográfica					
Aprendizaje de la metodología a seguir					
Descarga de los datos					
Tratamiento y preparación de los datos					

Construcción y entrenamiento de la red neuronal					
Establecimiento de conclusiones y comparación con la bibliografía recolectada					
Redacción de la memoria					
Entrega de la memoria					
Defensa del TFG					

4. Materiales y métodos

4.1. Búsqueda bibliográfica

Para la consulta de información sobre este trabajo se buscaron artículos en diferentes bases de datos y buscadores como ScienceDirect, Google Scholar o PubMed, además de diferentes páginas webs especializadas en programación y en concreto de ML. Para los datos propios para la realización de la red se obtuvieron de los Datos Abiertos de la Comunidad de Madrid y del Banco de Datos Municipal y Zonal de la Comunidad de Madrid.

4.2. Proceso experimental

La parte experimental de este trabajo se dividió en dos grandes tareas: la obtención a partir de la descarga de datos abiertos sobre los contagios en la Comunidad de Madrid y su posterior tratamiento y la construcción y entrenamiento de la red neuronal.

4.2.1. Obtención y filtrado de datos

Como se ha mencionado anteriormente, para realizar la parte experimental de este trabajo se recopilaron datos tanto del banco de Datos Abiertos de la Comunidad de

Madrid como del Banco de Datos Municipal y Zonal de la Comunidad de Madrid. En la primera se consultaron datos sobre los contagios, tasa de incidencia y municipio o distrito con la fecha correspondiente, siendo estos datos semanales desde el momento en el que se obtuvieron registros por municipio y distrito (3 de marzo de 2020) hasta el 4 de mayo de 2021.

Obtenidos estos datos, se trataron para adecuarlos al formato que requería la red: se filtraron los dos ficheros, ya que existía uno desde el 3 de marzo de 2020 hasta el 1 de julio de 2020 y otro desde esta última fecha hasta el 4 de mayo de 2021, de manera que únicamente quedaran las variables necesarias (fecha, contagios, incidencia y código del municipio).

Una vez construida la red, se introducirán datos para la predicción que serán exclusivos de los municipios de interés en función de determinados datos socioeconómicos, que en este caso serán los relativos a la renta per cápita. Estos ficheros se construyeron de la misma manera que se filtraron los anteriores en los cuáles se encontraban todos los municipios y distritos de la comunidad y únicamente se incluyeron las variables mencionadas anteriormente.

Por último, se descargaron múltiples ficheros sobre datos socioeconómicos acerca de los municipios y distritos de la Comunidad de Madrid para posteriormente conocer qué municipios son los que encabezan estos ránquines de interés y por el contrario quienes se encuentran a la cola.

Los datos para la construcción de la red los podremos considerar como una serie temporal, cosa que favorece al tipo de red escogida (Bagnato, 2019)

4.2.2. Construcción de la red neuronal

Una vez obtenidos y filtrados los datos necesarios para la construcción de la red, se procedió a la construcción de la misma.

Respecto a las redes neuronales, existen varios tipos entre los cuáles se eligió una arquitectura para la red de *feedforward* o también llamada perceptrón multicapa o MLP por sus siglas en inglés (*multi-layered perceptron*). Este tipo de red neuronal se caracteriza por que las salidas de las neuronas de la capa anterior se convierten en entradas para la siguiente sin necesidad de introducir más datos, además de que todas las neuronas de entrada están conectadas con todas las de la siguiente capa.

Este tipo de red es la ideal para predecir series temporales, ya que permite realizar predicciones (*forecasting*) a partir de un entrenamiento en el cual las convertiremos en un problema de tipo supervisado que permitirá utilizar el método de *backpropagation* (Bagnato, 2019).

Ha sido la utilizada por lo general para predecir la evolución de los casos y posibles nuevos focos en diferentes países, siendo una herramienta fundamental para el control de la pandemia (Tamang, Singh y Datta, 2020).

Con problema de tipo supervisado nos referimos a que la red tiene unas salidas más sencillas que se comprenden en un número reducido o incluso binario, de manera que para la red será más fácil de manejar y posteriormente podremos revertir estos valores a los reales (Hastie, Tibshirani y Friedman, 2009).

El término *backpropagation* o propagación hacia atrás es un método de aprendizaje que permite ajustar los pesos de las conexiones entre las neuronas de forma automática a medida que se entrena el modelo (Rumelhart, Hinton y Williams, 1986).

Previamente al comienzo de la creación de la red, se dispondrá de las herramientas necesarias para ello: el lenguaje de programación, en este caso Python; las librerías Keras, Tensorflow y Pandas. Python nos servirá como cualquier otro lenguaje de programación para ejecutar los códigos, pero fue escogido este por su sencillez. Respecto a las librerías, las dos primeras se escogieron por ser ideales para la construcción de redes neuronales: la primera sirve para ejecutar la segunda y está creada para funcionar con Python. Tensorflow es una de las librerías más utilizadas para la predicción a partir de series temporales. Para ejecutar y simplificar el uso de todas estas herramientas se utilizó el sistema de gestión de paquetes Anaconda, que es una distribución abierta y libre de los lenguajes de programación (Bagnato, 2019) Python y R. Pandas es una librería exclusiva de Python para tratamiento de datos. (McKinney, 2011).

Escogida la estructura de la red neuronal, se introducen los datos y el código elegido (Bagnato, 2019), utilizando las herramientas mencionadas anteriormente.

Tras introducir los datos, se tratan los datos cómo hemos explicado previamente: se transforma la serie temporal en un problema de tipo supervisado para poder entrenarla con el concepto de propagación hacia atrás.

En esta red se utilizó para entrenarla a partir del método de *backpropagation* las siete semanas previas para predecir una octava, es decir, se agrupan de 7 en 7 semanas y la semana siguiente es la que predice el modelo para entrenar.

Antes de ejecutar esta función, se transforman los valores entre -1 y 1 para que a la red le sea más sencillo realizar los cálculos. Posteriormente se transforman para observar los datos que queremos.

	var1(t-7)	var1(t-6)	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
7	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000
8	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648
9	-0.999648	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648	-1.000000
10	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000
11	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648	-1.000000	-1.000000	-0.999648

Figura 6-. Muestra de los datos transformados y preparados para introducirlos en la red. Las siete primeras columnas, es decir, de var1(t-7) hasta var1(t-1) serían entradas y la última columna var1(t) sería la salida.

En este punto los datos estarían listos para crear la red. Con anterioridad se debe elegir qué cantidad de datos irá destinada al conjunto de entrenamiento (del inglés *train*) y al conjunto de pruebas (del inglés *test*). El porcentaje más común suele ser 80% entrenamiento y 20 % pruebas, pero se pueden probar diferentes porcentajes dependiendo del número de datos disponible y de cuánto de compleja debe ser la red para estos. En este caso se utilizó este porcentaje, siendo en principio el estándar. Las entradas de datos totales eran de 12,310, siendo 8,728 del año 2020 y 3,582 del 2021. De estos datos, se reservaron para el conjunto de pruebas 2,469 entradas.

Una vez se obtienen los datos separados en conjunto de entrenamiento y conjunto de pruebas, se determina la arquitectura de la red determinando las capas y neuronas que tendrá cada uno de los tres tipos de capas: las entradas, las capas ocultas y las salidas.

Tras varias pruebas, la que aparentemente mejor resultados ha dado es la siguiente:

-Siete neuronas de entrada.

-Una capa oculta con siete neuronas.

- Una sola salida, que corresponderá al número de casos en cada municipio.

Como función de activación se utilizó la función de la tangente hiperbólica, la cual tendrá valores entre -1 y 1 y será la adecuada para esta red, ya que los valores estarán comprendidos entre estos dos números.

Para calcular el error y saber cómo de preciso es nuestro modelo para la predicción se aplicó el error cuadrático medio o MSE (del inglés *mean squared error*). Al tener valores continuos y no discretos, este error es el adecuado. A medida que el número de *epochs* aumente, este debería descender su valor.

Construidas todas las capas de la red, el siguiente paso es entrenar la red. Para ello, también de forma arbitraria y después de bastantes pruebas con diferentes valores, se entrenó la red con 750 *epochs*, intentando encontrar un valor que permita no sobreentrenar (del inglés *overfitting*) la red ni infrajustarla (del inglés *underfitting*). Este proceso es totalmente experimental y dependerá tanto de los datos como de la estructura de la red neuronal.

Finalizados los *epochs*, se puede observar cómo el MSE se reduce drásticamente del primer *epochs* al último, pasando de 0.0712 a 4.0993e-04, siendo prácticamente 0.

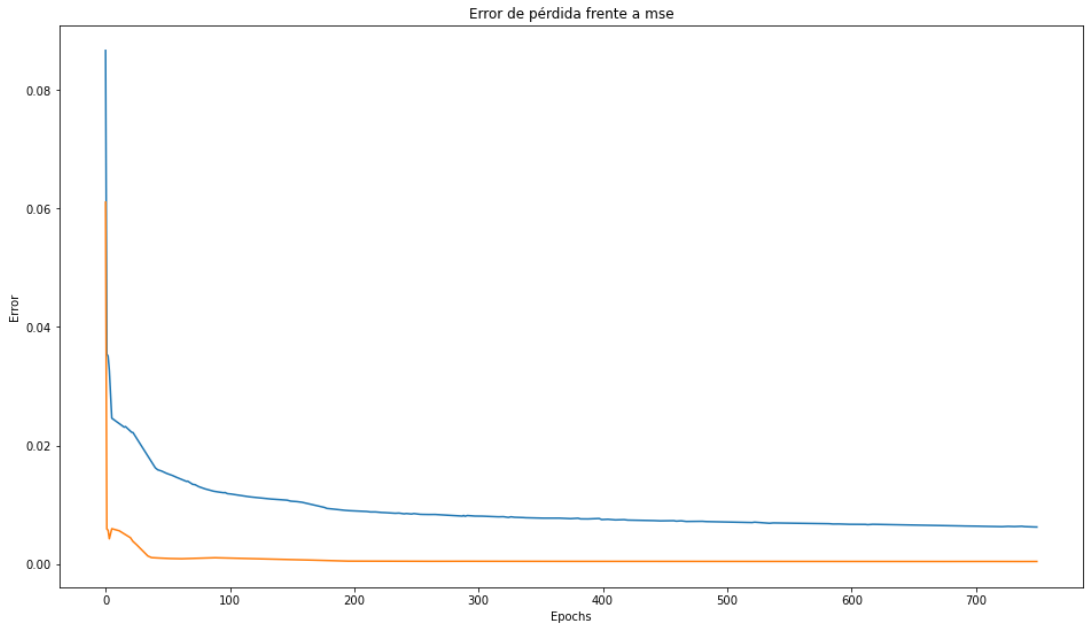

Figura 7-. Representación gráfica del error de entrenamiento (azul) y el error cuadrático medio (naranja).

Una vez la red está entrenada y los *epochs* han finalizado, podemos visualizar el conjunto de validación:

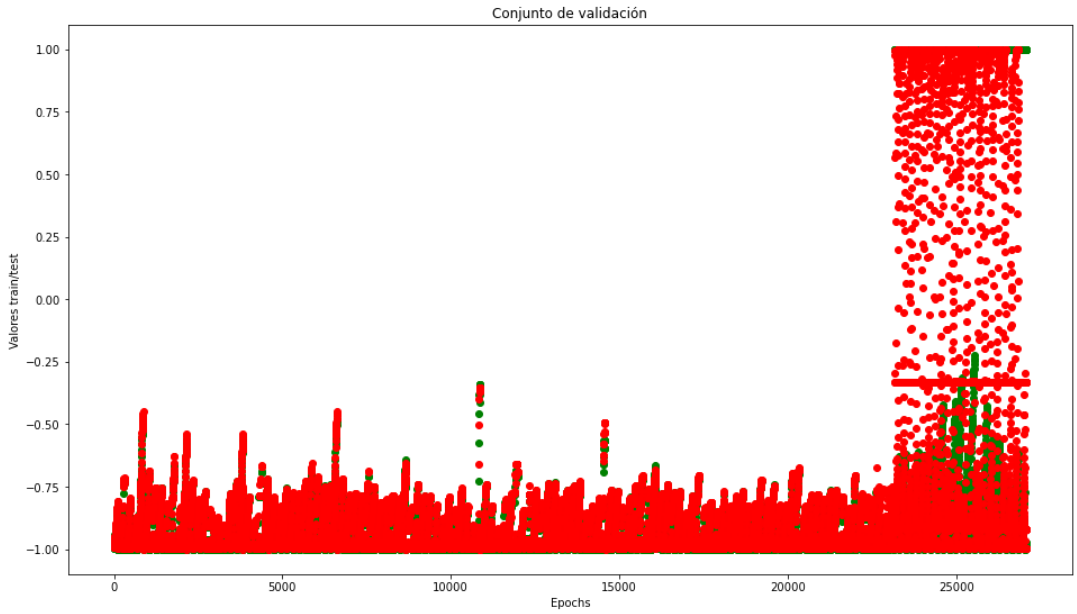

Figura 8-. Conjunto de validación de la red neuronal.

El conjunto de validación, que se consigue mediante la técnica de validación cruzada o en inglés *cross-validation*, ayuda a identificar si el modelo es el adecuado o no. Para ello, los puntos verdes deberán estar alineados lo máximo posible a los rojos.

El conjunto mejora por lo general cuanto más aumentamos los *epochs* hasta un punto en el que ya no puede. Entrenada la red, se puede proceder a las predicciones de los municipios de interés.

Finalmente, para realizar las predicciones y obtener los resultados se realiza un pronóstico añadiendo los datos de cada municipio por separado, de manera que estos municipios escogidos (los tres que más renta per cápita tienen frente a los que menos) serán los que tendrán predicciones de 3 semanas vista desde el 4 de mayo de 2021.los que menor renta per cápita neta media por persona tendrían con datos actualizados de 2017 serían Alcalá de Henares, Parla y Aranjuez con 6.082, 6.448 y 6.778 respectivamente; mientras que los otros tres municipios que tendrían los mayores valores de renta per cápita serían Alcobendas, Boadilla y Majadahonda con 30,210 euros cada uno.

El proceso para realizar la predicción no es muy diferente del anterior en lo que se refiere a tratar los datos. La principal diferencia es que la última columna, es decir, la columna de salida que antes era var1(t) ahora estará vacía por ser la columna que debe predecir el modelo. Predecirá las 3 semanas siguientes desde la última semana que hay registros (Bagnato, 2019).

5. Resultados

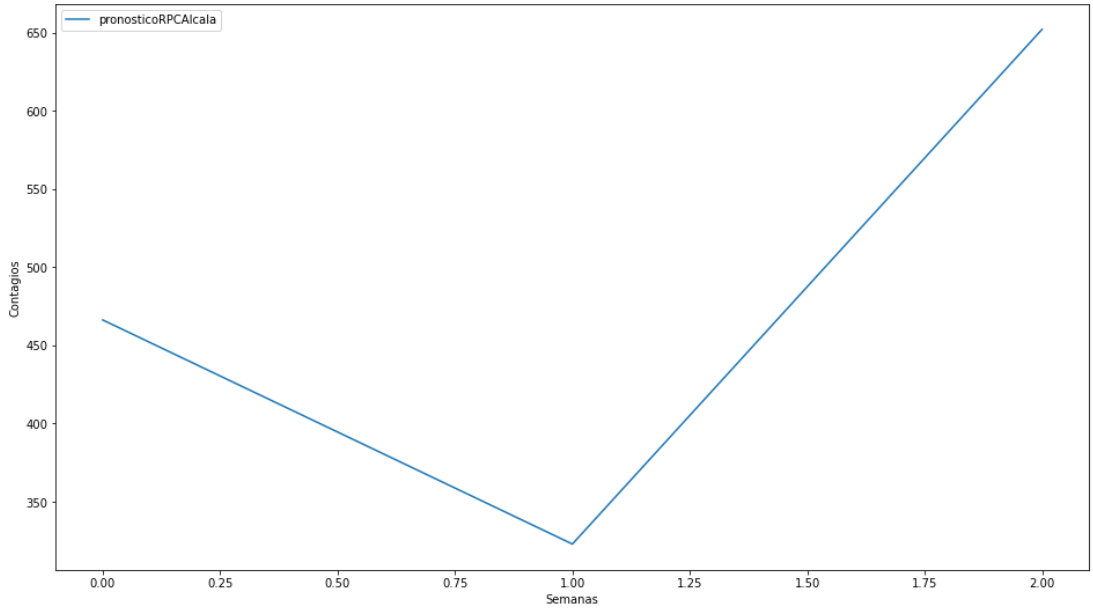

Figura 9-. Predicción de la evolución de los contagios a 3 semanas vista en el municipio de Alcalá de Henares. Los valores del pronóstico son los siguientes:466, 322 y 652 casos.

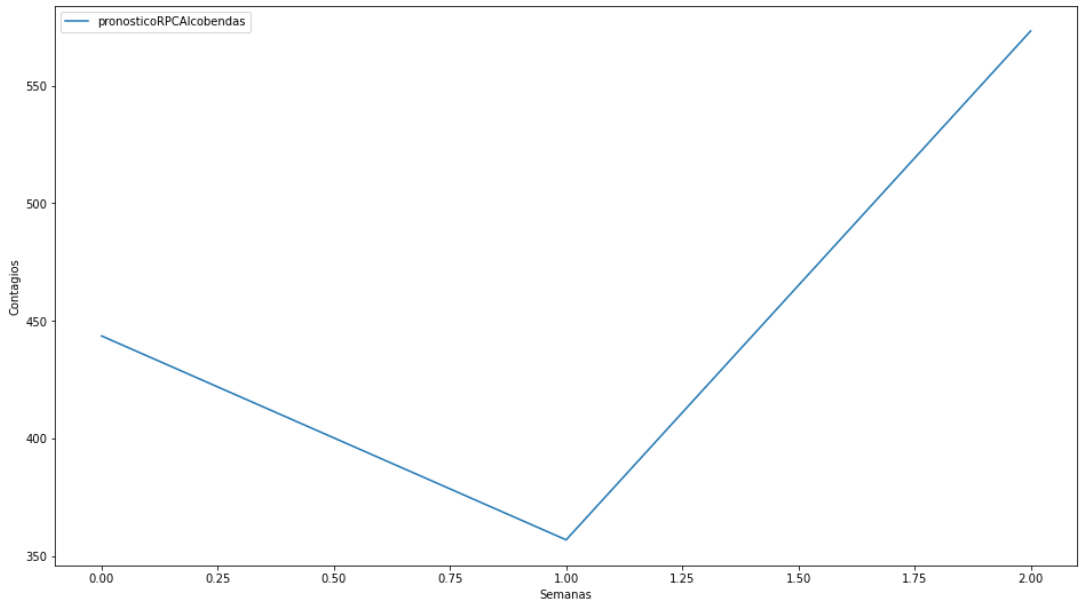

Figura 10-. Predicción de la evolución de los contagios a 3 semanas vista en el municipio de Alcobendas. Los valores del pronóstico son los siguientes:443, 356 y 573 casos.

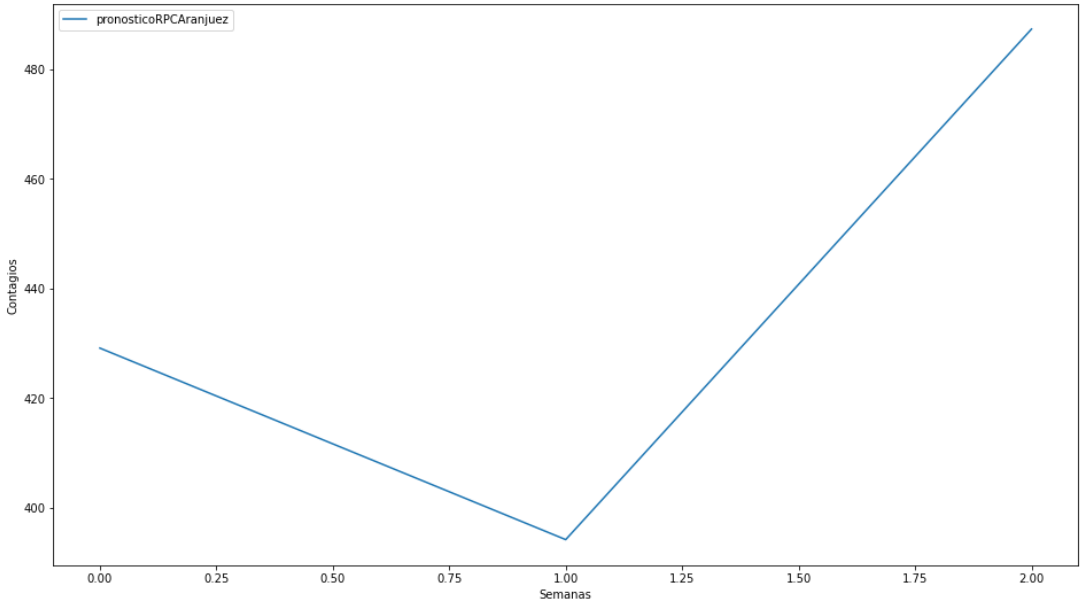

Figura 11-. Predicción de la evolución de los contagios a 3 semanas vista en el municipio de Aranjuez. Los valores del pronóstico son los siguientes:429, 394 y 487 casos.

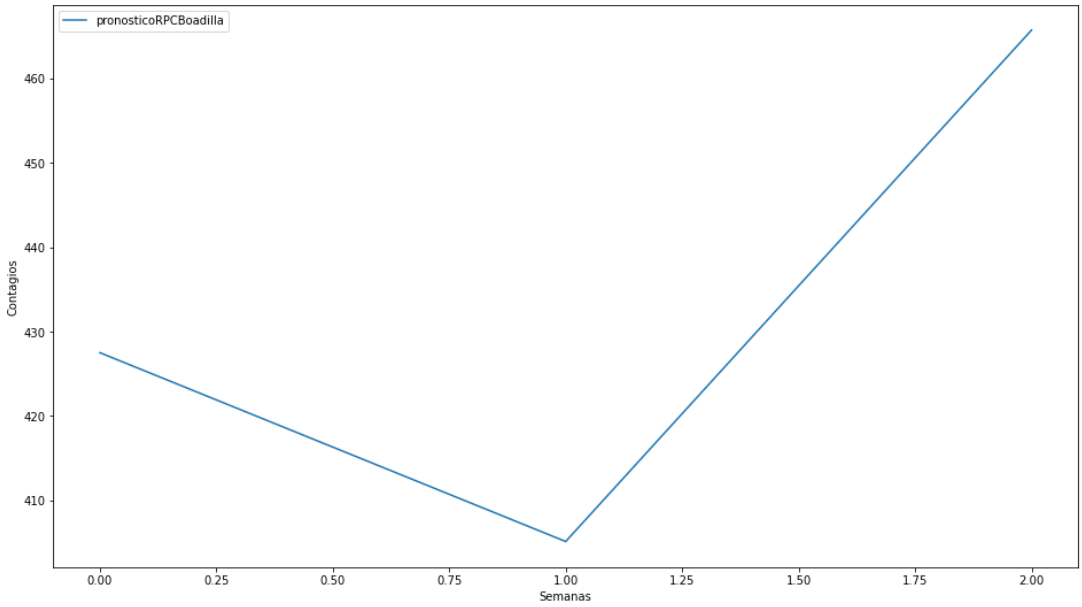

Figura 12-. Predicción de la evolución de los contagios a 3 semanas vista en el municipio de Boadilla. Los valores del pronóstico son los siguientes:427, 405 y 465 casos.

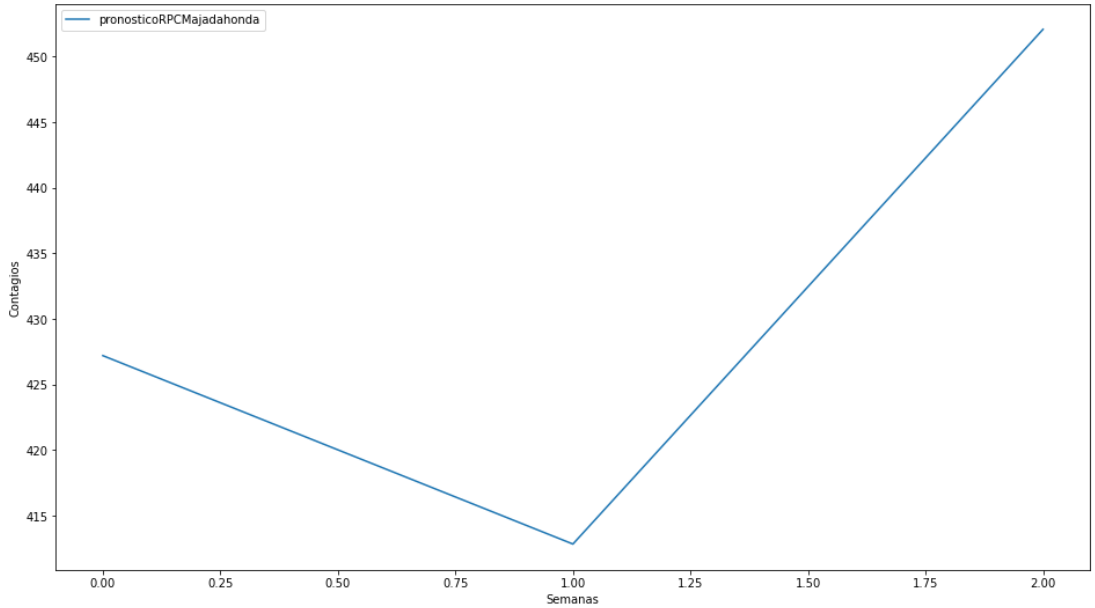

Figura 13-. Predicción de la evolución de los contagios a 3 semanas vista en el municipio de Majadahonda. Los valores del pronóstico son los siguientes: 427, 412 y 452 casos.

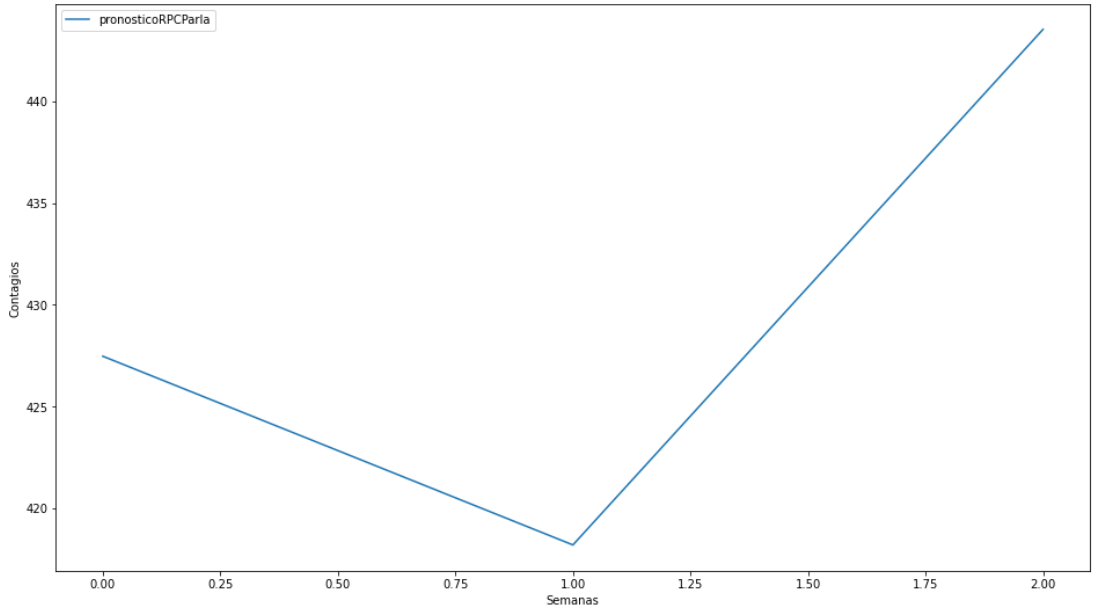

Figura 14-. Predicción de la evolución de los contagios a 3 semanas vista en el municipio de Parla. Los valores de pronóstico son los siguientes: 427, 418 y 453 casos.

6. Discusión

Para proceder a la discusión del estudio, se dividió también en dos partes: la discusión de la construcción de la red neuronal, que se basará en exponer mejoras a la red planteada a partir de redes neuronales similares y la discusión de los propios

resultados mediante bibliografía. En este último apartado se discute acerca de los factores que pueden acelerar o disminuir la propagación de la COVID-19, ya sean socioeconómicos o ambientales.

6.1. Discusión de la construcción de la red

La construcción de la red puede mejorarse de diferentes formas, exponiéndose las principales a continuación.

6.1.1. *Overfitting*

El sobreajuste u *overfitting* es un problema central en el aprendizaje automático supervisado y en concreto en las redes neuronales que en muchos casos dificulta la generalización de los modelos para que estos se ajusten correctamente a los datos de entrenamiento y a los datos de pruebas.

Para comprobar si existe sobreajuste en el modelo, es necesario realizar una gráfica en la cual se represente el error de validación y el error de pruebas. En los entrenamientos con redes neuronales existe un punto ideal en el que ni existe *overfitting* ni *underfitting*, de manera que el objetivo es encontrar este punto en la gráfica. Este punto existe en el momento en el cual el error de validación no aumenta ni desciende mientras el error de pruebas disminuye.

El sobreajuste puede producirse por diferentes causas, pero podemos clasificarlos en tres principales:

-Ruido en el conjunto de entrenamiento. Este fenómeno se produce cuando el conjunto de datos no es representativo o es excesivamente pequeño, lo que hace que este ruido pueda ser aprendido y no preparar al modelo para diferentes situaciones

-Complejidad de las entradas. Una gran variedad en las entradas que no permita a la red identificar determinados valores como representativos, el modelo se vuelve demasiado preciso y poco consistente.

-Múltiples procedimientos de comparación. En varias ocasiones, los criterios de selección propios de los algoritmos, como las funciones de activación o de coste, hacen selecciones que no mejoran el modelo o incluso reducirán su precisión (Ying, 2019).

Para este modelo, en principio las causas más problemas del *overfitting* presente en la red serían las dos primeras. Podría existir ruido que el modelo haya aprendido al añadir como entrada la etiqueta de identificación de cada municipio, pudiendo

introducir esta etiqueta fuera de la red y que únicamente fuese utilizada como identificador.

Para ello, sería necesario introducir en el programa una función de índice múltiple, siendo índices la fecha y la etiqueta de identificación del municipio (Bagnato, 2019).

La segunda causa del sobreajuste ha podido ser causada por la variedad y seguramente no tantas entradas para cada municipio. Los datos son semanales y hay unas 61 entradas por cada municipio, lo que a pesar de sumar más de 12,000 entradas en total para tantos municipios pueden ser pocos datos y demasiada variedad, aumentando en exceso la complejidad de las entradas.

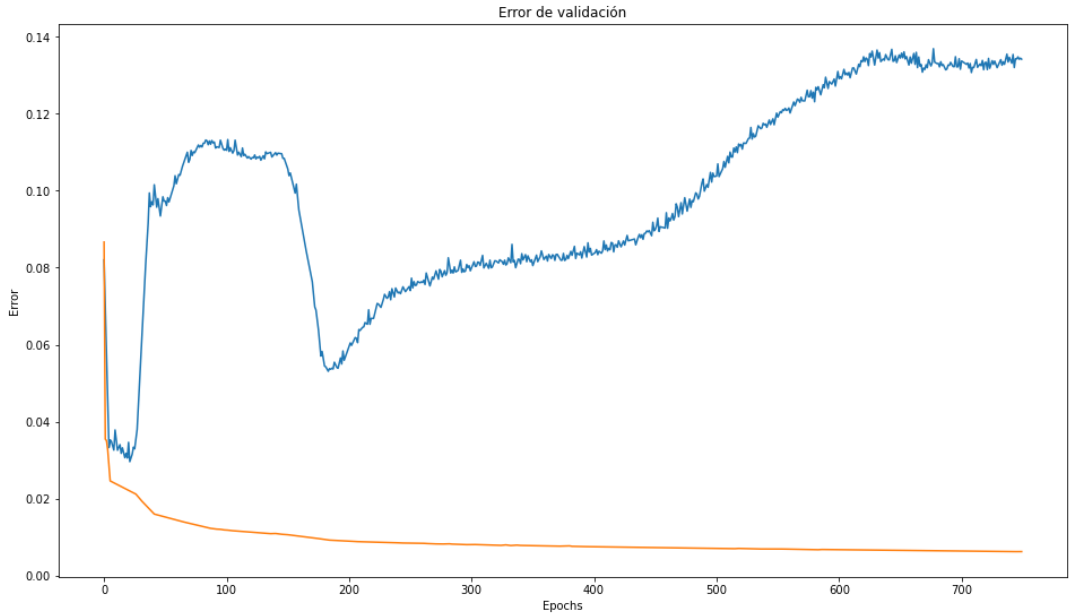

Figura 15-. Representación gráfica de la evolución del error de validación (azul) y del error de entrenamiento (naranja).

En este caso (Figura 15), existe un sobreajuste evidente. El error de validación sigue con ascensos y descensos mientras el error de pruebas continúa un descenso paulatino.

Las posibles soluciones para resolver el sobreajuste podrían ser las siguientes:

-Detener el entrenamiento un punto ideal en el que no exista ni *overfitting* ni *underfitting*.

-Reducción de la red. Esto significa reducir datos que puedan ser irrelevantes o no significativos para el entrenamiento de la red y reducir su complejidad.

-Aumentar el conjunto de datos de entrenamiento. En este caso podemos solucionarlo de dos formas: o bien aumentando el porcentaje del conjunto de datos de entrenamiento respecto al de pruebas, o lo más lógico a poder ser, añadir más datos a la red.

En esta última opción sería importante no aumentar la complejidad de las entradas (Ying, 2019).

Intentar parar el entrenamiento antes de llegar a ese punto (Figura 16) parece complicado y poco útil, ya que parece llegar a un punto con las características anteriormente mencionadas cerca de los 10 primeros *epochs*, cuando el número de ciclos en una red neuronal tiene que ser mucho mayor y más cercano al número que hemos usado (750) para el tamaño de los datos.

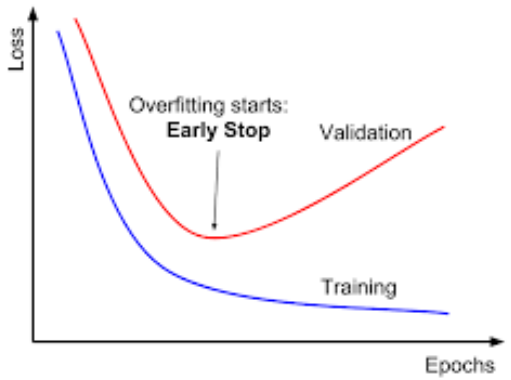

Figura 16-. Punto ideal para detener el entrenamiento y conseguir que el modelo no tenga ni overfitting ni underfitting. Es en este punto donde comienza el overfitting y ya no hay underfitting (Nuric, 2021).

6.1.2. Datos utilizados

Respecto a los datos utilizados, son entradas que quizás hayan tenido un exceso de complejidad o de datos representativos para los pocos datos que hay por municipio, de forma que seguramente se necesiten datos diarios para que la red sea lo suficientemente consistente y obtener así resultados fiables.

Son los datos que se extrajeron del Portal de Datos Abiertos de la Comunidad de Madrid, pero seguramente para realizar la red habría sido interesante aumentar la cantidad de datos de otra forma. Una manera podría haber sido aumentando el número de datos añadiendo datos semanales de más CCAA en la red, por ejemplo, empezando por las limítrofes, en las que exista una relación más estrecha por motivos económicos o culturales o por CCAA con una demografía u organización urbanística similar.

Finalmente se podría experimentar también introduciendo datos no sólo de España, sino también de lugares con características semejantes a la Comunidad de Madrid que sean relevantes para el modelo, es decir, que puedan influenciar en la evolución de los contagios.

6.2. Discusión de los resultados

Los resultados propios de la red neuronal no son del todo fiables por varios motivos. Además de los motivos propios de la construcción de la red explicados anteriormente, al crear el modelo e introducir los datos, la red no discriminó entre municipios, de forma que la red fue entrenada sabiendo el ID de cada municipio, pero sin distinguir entre ellos, lo que hace que las predicciones sean semejantes entre ellas.

Otra causa del error en la predicción, que complementa al anterior, es escoger los datos con los casos totales y no con los casos por días. La idea inicial era predecir a partir de estos, lo que proporcionaría unas predicciones que en ningún caso se asemejarían a las de los resultados, ya que existe un descenso de los casos en todos los municipios en la segunda semana que si la red hubiese interpretado como totales los casos jamás descendería. Con este concepto y en caso de que la red hubiese interpretado los datos de esta manera, la gráfica que se observaría sería bien diferente al observarse únicamente ascensos de casos en el caso de que existiesen.

Para analizar los datos se podría haber utilizado diferentes técnicas, como calcular el porcentaje que ha aumentado entre un periodo de tiempo de interés o el aumento de la pendiente de la gráfica.

6.2.1. Un caso práctico

Un ejemplo muy semejante al propuesto en este trabajo en cuanto a metodología y objetivos (Wieczorek, Siłka y Woźniak, 2020) también tiene como objetivo el desarrollo de una red neuronal para la predicción de la propagación de la COVID-19 utilizando exclusivamente el número de contagios diarios para la construcción de la red, además de un identificador geográfico con las coordenadas y el nombre de la región, que se eliminan para la construcción de la red.

Evita la sobrealimentación de datos alimentando a la red con 30 regiones por archivo, de manera que no exista un exceso de complejidad en las entradas en la red (Ying, 2019). En el caso de los días de predicción se escogieron 14, a pesar de no ser los que menor error reportaban, que eran 6. Esto sucede porque con un menor número de días para predecir la red se adapta mejor, pero a la vez la predicción es menos estable y puede ser más fluctuante. Se observó que al aumentar los días, a pesar de que aumentara el error, este era mínimo y los resultados necesitaban ser estables en el tiempo. En este trabajo se escogió sabiendo de antemano que existe un mayor error al aumentar los días

de predicción. Cuantos menos días, la red se adapta más rápido a los cambios y se comporta mejor a cambios de tendencia (Bagnato, 2019). Para los métodos de elección se realizan procesos de prueba; decidiendo de forma arbitraria el conjunto de datos, los días de predicción o el número de *epochs* necesarios para no tener *overfitting* ni *underfitting*, como se describió anteriormente.

Una mejora en el entrenamiento es la división de los datos, concepto válido para este trabajo: se dividen los datos sin excluir países de los conjuntos de entrenamiento y de prueba. De ser así, el modelo perdería precisión al entrenar con unas tendencias que son diferentes y podrían alterar la predicción. En los resultados obtenidos, la red es entrenada sin dividir los archivos previamente para que en ambos conjuntos existan datos de todos los municipios.

Se demostró que las redes neuronales tienen una mayor precisión y ajuste que los métodos estocásticos. También durante el proceso de investigación para encontrar el mejor método de predicción se utilizó una red neuronal recurrente (RNN), pero los resultados fueron de un 1 a un 2% menos precisos que con la red neuronal artificial clásica (ANN) utilizando el concepto de *backpropagation.* Además, consigue llegar al número de *epochs* ideal, en el cual alcanza la máxima precisión, en un número mucho menor que en la RNN.

Las principales desventajas de este modelo residen en el tiempo de entrenamiento de la red y el ajuste de los parámetros de la red. Con este modelo, se consiguió una precisión media para todos los países de 87.70%, suficiente para medir la tendencia de los contagios. Se añade que se podrían utilizar predictores individuales para cada país o región en un futuro para mejorar las predicciones.

Esta investigación también muestra dos puntos que dificultaron y/o dificultan el trabajo de la predicción de los contagios de la COVID-19: la falta de datos al comienzo o incluso ahora de los contagios en algunos países o regiones y el comportamiento de la población, el gobierno o el acceso a información acerca de la enfermedad y equipos médicos. Esto último expone una cuestión fundamental en la evolución de la pandemia: los factores sociales, económicos y el humano (Wieczorek, Siłka y Woźniak, 2020).

6.2.2. Factores que pueden determinar la evolución de la COVID-19

Existen múltiples factores que pueden influir en la propagación de la COVID-19 y es fundamental realizar estudios para profundizar acerca de cuáles pueden ser estos y actuar para mitigarlos o intentar que influyan lo menos posible. Para estudiar cuáles

podrían ser estos factores en Brasil, se estudiaron 14 factores diferentes a partir de una red neuronal como la creada en este trabajo, ANN, del tipo SOM (mapas autoorganizados o en inglés *self-organizing maps*) y así separar a cada unidad federativa brasileña y estudiarlas de forma individual. Tras el estudio se determinó que factores como la vacunación antigripal, camas de unidad de cuidados intensivos (UCI), ventiladores, personal sanitario o el Índice de Desarrollo Humano (IDH) pueden ser determinantes para la propagación de la COVID-19. Se demostró que en las unidades federativas en las que existía una mayor vacunación antigripal, mayor número de camas de UCI, personales sanitario o IDH existe un menor número de contagios y muertes por la COVID-19. Sin embargo, otros factores que aparentemente podrían ser también determinantes para la propagación de la enfermedad como la cantidad de equipos de protección individual (EPIs), pruebas de detección del SARS-CoV-2, medicamentos para el tratamiento de la enfermedad o fondos federales no fueron determinantes para la evolución de la pandemia en Brasil (Galván *et al.*, 2020).

En la India, en un estudio con objetivos similares y utilizando una metodología diferente a las redes neuronales, se determinó que la propagación de la COVID-19 es dependiente de la presión, humedad relativa, temperatura y velocidad del viento; con un 76.08% de dependencia de casos en estos factores (Jha *et al*., 2021).

En el caso de la Comunidad de Madrid, se ha demostrado que existe una correlación negativa entre el número de casos y la distancia a Madrid significativa, es decir, a más distancia de la capital, menos casos de COVID-19. La variable estudiada en este trabajo, la renta per cápita, es la única variable socioeconómica que demuestra tener un coeficiente estadístico significativo, siendo la propagación de casos de la enfermedad dependiente de este factor. Cerca del 67% de los casos son dependientes de dos factores: la temperatura y la densidad poblacional, siendo la distancia a Madrid no tan relevante. Esto sugieren los datos de contagios que ofrece la Comunidad de Madrid, donde se observa un descenso de los casos durante el comienzo de la época estival (Oto-Peralías, 2020).

Sin embargo, en un estudio que trata el inicio de la pandemia en la Comunidad de Madrid, la propagación de la COVID-19 no está relacionada con la densidad poblacional, al menos en los distritos de la ciudad de Madrid. Una de las zonas más densamente pobladas y más complejas urbanísticamente de la ciudad de Madrid como la "almendra central" no fue un lugar de concentración de los casos de COVID-19, siendo en zonas periféricas donde se detectaron más casos.

También se demuestra que la renta per cápita sí es una variable que influye en la evolución de los casos de COVID-19. Se probó que los distritos más vulnerables y con menos renta per cápita de la ciudad, con cerca de 12,000 euros menos que los más ricos, existía un mayor aumento de casos y muertes. Estos distritos con las rentas per cápita más bajas de la ciudad son las que presentan también una mayor presencia de factores que aumentan la mortalidad por COVID-19, como el sexo (hombre), la edad (mayores de 65 años), y comorbilidades presentes en la población como hipertensión, diabetes, enfermedades cardiovasculares y accidentes cerebrovasculares (Menéndez e Higueras, 2020).

Otro factor determinante que se ha demostrado que favorece a la propagación de la COVID-19 de forma experimental es la contaminación. Este factor provoca una mayor propagación del virus, ya que la contaminación puede ser un transmisor incluso más eficaz que una persona contagiada (Coccia, 2020). Los niveles de Madrid, especialmente en la periferia del centro de la capital y al sureste, son especialmente elevados. La ciudad de Madrid supera con creces los niveles de contaminantes en el aire recomendados por la OMS (Núñez-Alonso,2019; Menéndez e Higueras, 2020).

6.2.3. Ejemplos similares en otras grandes ciudades

Por ejemplo, en China, Wuhan, también se estudiaron los posibles factores socioeconómicos que podrían determinar la evolución de la pandemia. Como en el estudio sobre la Comunidad de Madrid acerca de estos factores, en la capital de la provincia de Hubei se detectó una correlación positiva tras la reducción de las restricciones entre la renta per cápita y los contagios, seguramente asociada a un aumento de la actividad económica y social (Qiu, Chen y Shi, 2020).

El ejemplo de Wuhan puede resultar una excepción. En Europa, seguramente un ejemplo más representativo para nuestro objeto de estudio, durante la primera ola, que abarcaría del 1 de abril al 31 de mayo, se observó que los países con una renta per cápita mayor sufrieron un cambio menor en el número de casos. También se demostró que el tabaquismo en la población está correlacionado positivamente con los casos de COVID-19 y que el hecho de que en los países con mayor renta per cápita exista un cambio menor en el número de casos sugiere también que los déficits de inversión en la salud pública son un problema para la población de los países con una renta per cápita menor, como es el caso de Bulgaria, Ucrania o Rumania. El pobre acceso a la sanidad en estos países por gran parte de la población puede ser determinante para su salud, siendo en

tiempos de crisis cuando estas dificultades se agravan y enfermedades que antes no eran un problema serio para los pacientes acabaron por provocar complicaciones en enfermedades o problemas de salud que previos a la pandemia eran residuales. La disminución en la prestación de servicios de salud provocó un descenso de más del 50% en la derivación urgente de pacientes, esperando incluso 2 semanas pacientes con un posible diagnóstico de cáncer. En esta situación, en los lugares con más dificultades económicas la situación aún puede ser más grave: en las zonas más pobres el trabajo manual es más necesario, donde no hay tantas opciones en el mercado laboral; mientras que en zonas con un mayor poder adquisitivo el mercado laboral ofrece oportunidades que en el contexto vivido permiten seguir desarrollando sus tareas laborales desde casa mediante el teletrabajo, lo que sobre todo al inicio de la pandemia, en una situación de confinamiento domiciliario o restricciones horarias puede reducir la propagación del virus (Neal, 2020;Pardhan y Drydakis, 2020).

En la Comunidad de Madrid se puede extrapolar un estudio como este a sus barrios y municipios más empobrecidos o con un menor acceso a los servicios de la salud pública, como los del sureste de la capital, donde se concentraron más los casos.

De la misma forma, estas zonas tenían una red de apoyo mayor para paliar estas dificultades de la población tanto en el acceso a la sanidad pública como a alimentos o servicios de primera necesidad (Menéndez e Higueras, 2020).

7. Conclusiones

En este trabajo se ha intentado crear una red neuronal capaz de predecir los contagios de la enfermedad COVID-19 a partir de datos de contagios de toda la Comunidad de Madrid para predecir en diferentes municipios la evolución de la pandemia. A pesar de no haber conseguido resultados fiables, se ha expuesto de forma crítica y razonada todas las mejoras posibles del modelo para un futuro perfeccionamiento del mismo a partir de otros modelos más robustos, en los cuales se exponen limitaciones intrínsecas del modelo realizado.

Atendiendo a los resultados obtenidos y la discusión realizada, se puede asegurar que existe una amplia variedad de factores que determinan la propagación de la COVID-19 en todo el mundo. Estos factores pueden llegar a ser muy diferentes entre sí e incluso en algunos lugares y situaciones tener una correlación contradictoria. Por ello, es importante dirimir en qué contextos estos factores serán perjudiciales o beneficiosos para la evolución de la pandemia y actuar en consecuencia para reducir la propagación

de la enfermedad en la población. En nuevas "olas" de casos, sería fundamental que los gobiernos competentes de la gestión de sus territorios tomasen parte y actuasen con toda la información disponible y siempre teniendo en cuenta los datos e investigaciones necesarias para tomar las decisiones pertinentes para reducir la expansión de la COVID-19.

En relación a este estudio, a través de una búsqueda bibliográfica se ha podido asegurar cómo determinados factores socioeconómicos como la renta per cápita sí determinan la propagación de la pandemia en la Comunidad de Madrid. En cambio, otros como la densidad poblacional, un factor que se ha tenido en cuenta para decidir las restricciones en varias regiones, no han sido significativamente determinantes para la propagación de la COVID-19.

7.1. Limitaciones de los estudios sobre COVID-19

Las limitaciones que presentan los estudios sobre COVID-19 que necesitan de los datos de contagios supone una dificultad para cualquier investigación, asumiendo que existirán sesgos que en cada país o región que serán diferentes. Por ejemplo, el ritmo de pruebas de detección de la enfermedad en cada región puede ser diferente y no reflejar la situación real del área de estudio. Además, puede existir también dificultades para conseguir los datos necesarios para llevar a cabo un estudio, no solo de los contagios de COVID-19, sino también de otros sociológicos o económicos. En investigaciones que se centren en los contagios al inicio de la pandemia puede resultar un inconveniente añadido, ya que hasta que se pueda asegurar que el ritmo en la realización de pruebas de detección de la COVID-19 sea relativamente homogéneo en las áreas de estudio puede tardar un tiempo (Pardhan y Drydakis, 2020).

7.2. Futuras investigaciones

Para futuros trabajos de investigación, la aplicación de redes neuronales sería una herramienta fundamental no sólo para la predicción de la propagación de la enfermedad, sino también para múltiples campos como la detección de la enfermedad en radiografías o imágenes clínicas (Bassi y Attux, 2021) o la predicción de otras variables de interés durante la pandemia como el número de muertes o la ocupación en camas UCI (Dhamodharavadhani, Rathipriya y Chatterjee, 2020).

En la actualidad, la dificultad para encontrar datos de calidad no parece un problema salvo en casos excepcionales. Teniendo unas bases de datos lo suficientemente amplias

y representativas para prácticamente cualquier área o región, estas permiten realizar pronósticos robustos y aumentar el riesgo en las predicciones añadiendo más días a pronosticar o ajustando el modelo con áreas geográficas de menor tamaño.

7. Conclusions

In this work an attempt has been made to create a neural network capable of predicting COVID-19 disease infections from infection data from the entire Community of Madrid in order to predict the evolution of the pandemic in different municipalities. In spite of not having obtained reliable results, all the possible improvements of the model have been exposed in a critical and reasoned way for a future improvement of the same one from other more robust models, in which intrinsic limitations of the realized model are exposed.

Based on the results obtained and the discussion, it can be said that there is a wide variety of factors that determine the spread of COVID-19 throughout the world. These factors may be very different from each other and in some places and situations may even have a contradictory correlation. It is therefore important to determine in which contexts these factors will be detrimental or beneficial to the evolution of the pandemic and to act accordingly to reduce the spread of the disease in the population. In new "waves" of cases, it would be essential that the governments responsible for the management of their territories take part and act with all the information available and always taking into account the data and research necessary to make the relevant decisions to reduce the spread of COVID-19.

In relation to this study, through a bibliographic search it has been possible to ascertain how certain socioeconomic factors such as per capita income do determine the spread of the pandemic in the Community of Madrid. On the other hand, other factors such as population density, a factor that has been taken into account in deciding the restrictions in several regions, have not been significantly determinant for the spread of COVID-19.

7.1. Limitations of COVID-19 studies

The limitations of COVID-19 studies that require data on infection pose a difficulty for any research, assuming that there will be biases in each country or region that will be different. For example, the rate of testing for the disease in each region may be different and not reflect the real situation in the study area. In addition, there may also

be difficulties in obtaining the necessary data to conduct a study, not only of COVID-19 infections, but also of sociological or economic ones. In research that focuses on infections at the onset of the pandemic, this may be an added drawback, as it may take some time to ensure that the pace of COVID-19 testing is relatively homogeneous across study areas (Pardhan and Drydakis, 2020).

7.2. Future research

For future research work, the application of neural networks would be a fundamental tool not only for the prediction of the spread of the disease, but also for multiple fields such as the detection of the disease in radiographs or clinical images (Bassi and Attux, 2021) or the prediction of other variables of interest during the pandemic such as the number of deaths or occupancy in ICU beds (Dhamodharavadhani, Rathipriya and Chatterjee, 2020).

At present, the difficulty in finding quality data does not seem to be a problem except in exceptional cases. With sufficiently large and representative databases for practically any area or region, these allow robust forecasts to be made and increase the risk in the predictions by adding more days to forecast or adjusting the model with smaller geographic areas.

8. Bibliografía

- Bagnato, J.I. (2019). Pronóstico de series temporales com redes neuronales com Python (https://www.aprendemachinelearning.com/pronostico-de-series-temporales-con-redes-neuronales-en-python/)(Consultado el 3 de marzo de 2021).
- Bassi, P. R., & Attux, R. (2021). A deep convolutional neural network for COVID-19 detection using chest X-rays. *Research on Biomedical Engineering*, 1-10.
- Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology*, 188(12), 2222-2239.
- Burke, R. M., Killerby, M. E., Newton, S., Ashworth, C. E., Berns, A. L., Brennan, S., *et al*. (2020). Symptom profiles of a convenience sample of patients with COVID-19—United States, January–April 2020. *Morbidity and Mortality Weekly Report*, 69(28), 904.
- Chih-Cheng Lai, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, Po-Ren Hsueh(2020),Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents* 55:105924-3.
- Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., & Bernardini, S. (2020). The COVID-19 pandemic. *Critical reviews in clinical laboratory sciences*, *57*(6), 365-388.
- Coccia, M. (2020). Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. *Science of the Total Environment*, 729, 138474.
- Comunidad de Madrid (2020).(http://www.madrid.org/desvan/AccionLlamadaArbolDesvan_dwr.icm?tipoArbol=almudena) (Consultado el 15 de marzo de 2021).
- Condes, E., & Arribas, J. R. (2020). Impact of COVID-19 on Madrid hospital system. *Enfermedades Infecciosas y Microbiología Clínica*, 39(5), 256-257.
Datos contagios Covid-19 Comunidad de Madrid

- Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4), 197-387.
- Dhamodharavadhani, S., Rathipriya, R., & Chatterjee, J. M. (2020). COVID-19 mortality rate prediction for India using statistical neural network models. *Frontiers in Public Health*, 8.
- Ding, S., Li, H., Su, C. *et al*. (2013). Evolutionary artificial neural networks: a review. *Artif Intell Rev*, 39, 251–260.
- El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *machine learning in radiation oncology*, 3-11.
  EPData
- Estrada, P. (2016). Programando una red neuronal (https://bitybyte.github.io/Programando-una-red-neuronal/)(Consultado el 10 de julio de 2021).
- Fan, J., Ma, C., & Zhong, Y. (2019). A selective overview of deep learning. *arXiv*.
- Figueiredo, A. M., Daponte-Codina, A., Figueiredo, D. C. M. M., Vianna, R. P. T., de Lima, K. C., & Gil-García, E. (2020). Factores asociados a la incidencia y la mortalidad por COVID-19 en las comunidades autónomas. *Gaceta Sanitaria*.
- Galvan, D., Effting, L., Cremasco, H., & Adam Conte-Junior, C. (2020). Can socioeconomic, health, and safety data explain the spread of COVID-19 outbreak on Brazilian federative units?. *International journal of environmental research and public health*, 17(23), 8921.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In The elements of statistical learning. *Springer*, 9-41.
- https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos
- https://www.epdata.es/datos/evolucion-coronavirus-cada-comunidad/518/madrid/304
- Hu, B., Guo, H., Zhou, P., & Shi, Z. L. (2020). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, 1-14.
- INE (2020). (https://www.ine.es/jaxiT3/Datos.htm?t=2915)(Consultado el 29 de marzo de 2021).

- Jha, S., Goyal, M. K., Gupta, B., & Gupta, A. K. (2021). A novel analysis of COVID 19 risk in India incorporating climatic and socioeconomic Factors. *Technological forecasting and social change*, 167, 120679.
- Kukreja, H., Bharath, N., Siddesh, C. S., & Kuldeep, S. (2016). An introduction to artificial neural network. *Int J Adv Res Innov Ideas Educ*, 1, 27-30.
- McKinney, W. (2011). Pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9), 1-9.
- Menéndez, E. P., & Higueras García, E. (2020). Urban Sustainability Versus the Impact of Covid-19: A Madrid Case Study. *DisP-The Planning Review*, 56(4), 64-81.
- Mishra, M., & Srivastava, M. (2014). A view of artificial neural network. In *2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014)*, 1-3.
- Neal, K. (2020). The collateral damage of COVID-19. *J Public Health*, 42(4), 659.
- Núñez-Alonso, D., Pérez-Arribas, L. V., Manzoor, S., & Cáceres, J. O. (2019). Statistical tools for air pollution assessment: multivariate and spatial analysis studies in the Madrid region. *Journal of analytical methods in chemistry*.
- NURIC (2021). Imperial College Machine Learning – Neural Networks (https://www.doc.ic.ac.uk/~nuric/teaching/imperial-college-machine-learning-neural-networks.html)(Consultado el 12 de julio de 2021).
- Oto-Peralías, D. (2020). Regional correlations of COVID-19 in Spain. *OSF Preprints.*
- Pardhan, S., & Drydakis, N. (2020). Associating the change in new COVID-19 cases to GDP per capita in 38 European countries in the first wave of the pandemic. *Frontiers in Public Health*, 8, 1065.
- Qiu, Y., Chen, X., & Shi, W. (2020). Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. *Journal of Population Economics*, 33(4), 1127-1172.

- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.
- Sharif-Yakan, A., & Kanj, S. S. (2014). Emergence of MERS-CoV in the Middle East: origins, transmission, treatment, and perspectives. *PLoS pathogens*, 10(12), e1004457.
- Sharma, N., Sharma, R., & Jindal, N. (2021). Machine Learning and Deep Learning Applications-A Vision. *Global Transitions Proceedings*, *2*(1), 24-28.
- Tamang, S. K., Singh, P. D., & Datta, B. (2020). Forecasting of Covid-19 cases based on prediction using artificial neural network curve fitting technique. *Global Journal of Environmental Science and Management*, 6(Special Issue (Covid-19)), 53-64.
- Torres, J. (2018). Deep Learning: Introducción práctica con Keras. (https://torres.ai/deep-learning-inteligencia-artificial-keras/)(Consultado el 10 de julio de 2021).
- Uddin, M., Mustafa, F., Rizvi, T. A., Loney, T., Suwaidi, H. A., Al-Marzouqi, A. H. H. *et al.* (2020). SARSCoV-2/COVID-19: viral genomics, epidemiology, vaccines, and therapeutic interventions. *Viruses* 12:526.
- WHO: World Health Organization. (https://covid19.who.int/) (Consultado el 12 de julio de 2021)
- Wieczorek, M., Siłka, J., & Woźniak, M. (2020). Neural network powered COVID-19 spread forecasting model. *Chaos, Solitons & Fractals*, 140, 110203.
- Ying, X. (2019, February). An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, 1168(2), 022022.

9. Lista de Abreviaturas

OMS: Organización Mundial de la Salud

CCAA: Comunidades Autónomas

IA: Inteligencia artificial

SNN: Redes Neuronales de Impulsos (del inglés *spiking neural network*)

NN: Red neuronal (del inglés *neural network*)

ML: Machine Learning

10. Anexo

10.1. Anexo de los datos

Los datos son todos de código abierto y se descargaron de las siguientes páginas web:

Datos de los contagios en los municipios y distritos de la Comunidad de Madrid:

https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos

Datos socioeconómicos sobre los municipios de la Comunidad de Madrid:

http://www.madrid.org/desvan/AccionLlamadaArbolDesvan_dwr.icm?tipoArbol=almudena

10.2. Anexo de figuras adicionales

Las figuras fueron consultadas en la siguiente página web:

https://www.epdata.es/datos/evolucion-coronavirus-cada-comunidad/518/madrid/304

Estas figuras sirven para tener una perspectiva de la situación de la propagación de la COVID-19 en España por CCAA.

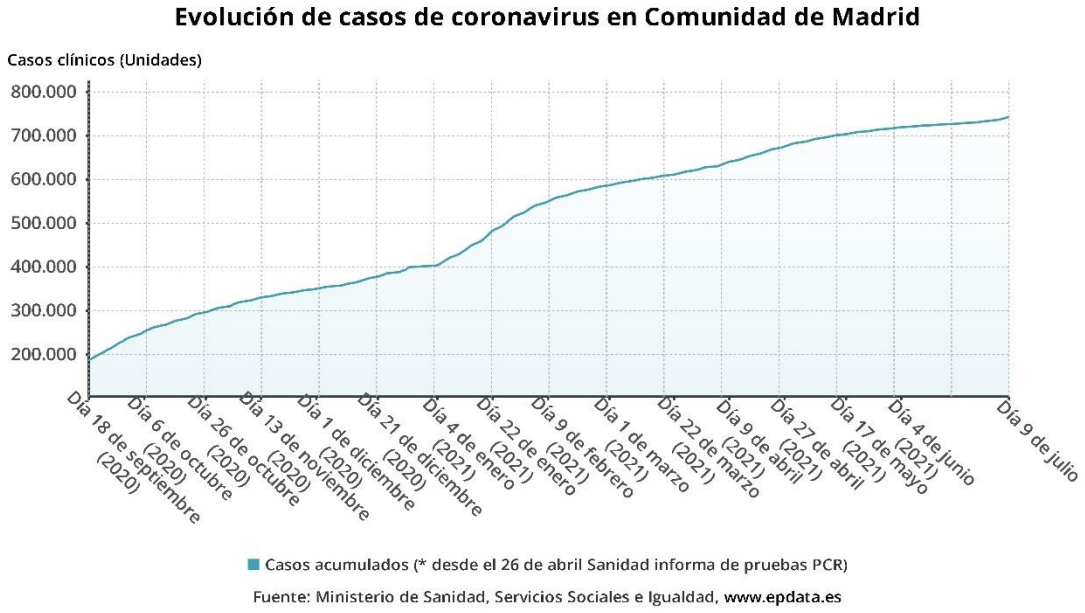

Figura 17-. Casos totales acumulados en la Comunidad de Madrid desde el 18 de septiembre de 2020.

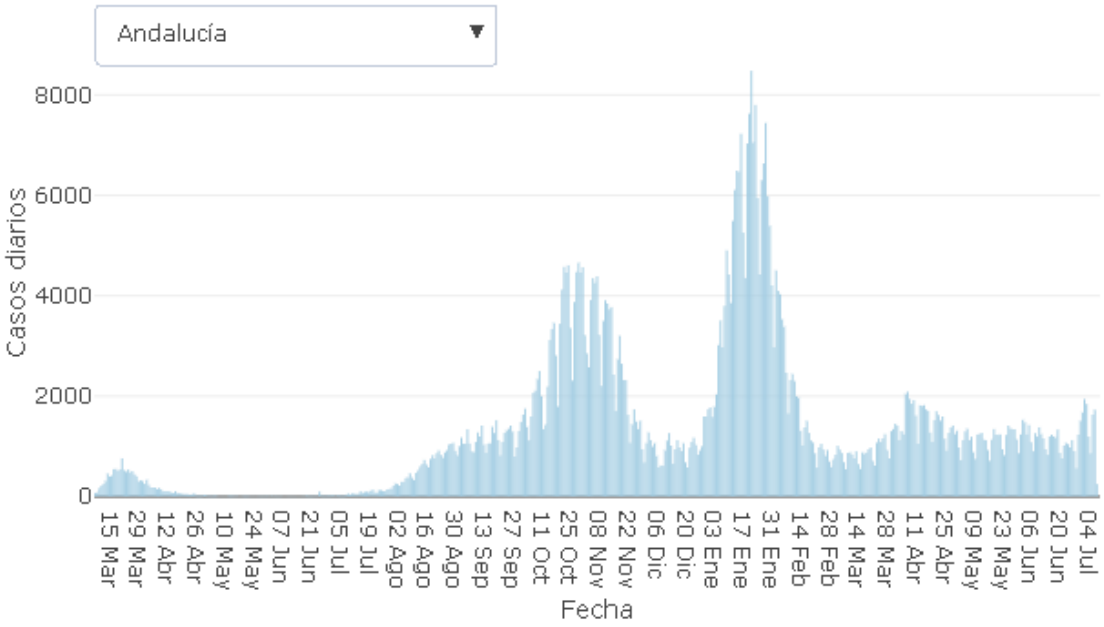

Figura 18-. Contagios diarios totales en Andalucía desde que se tiene registros (15 de marzo) hasta la actualidad.

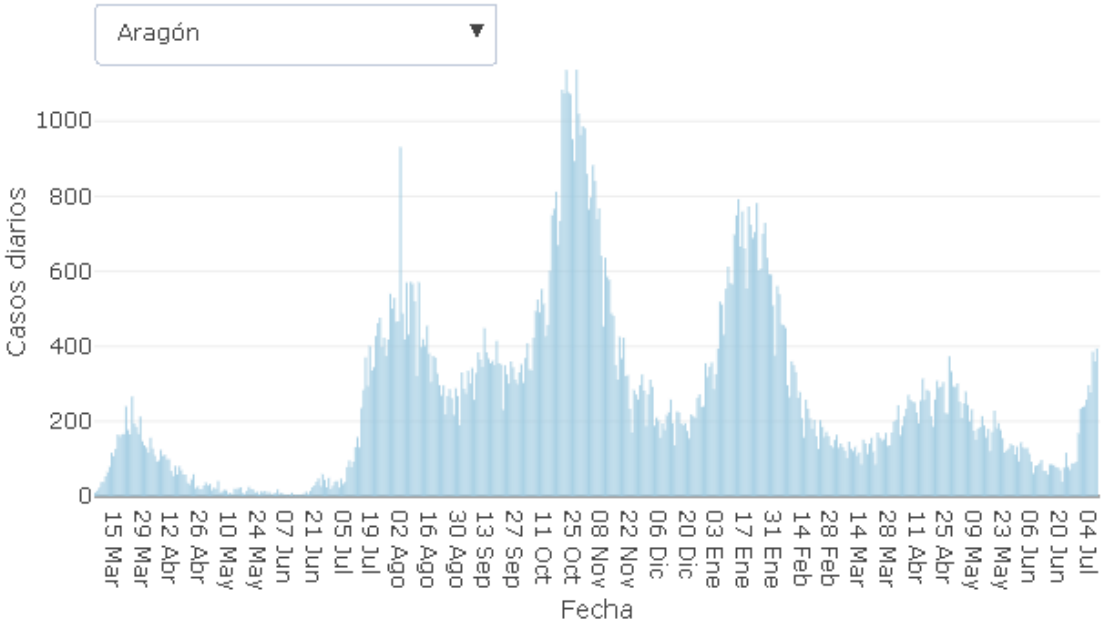

Figura 19-. Contagios diarios totales en Aragón desde que se tiene registros (15 de marzo) hasta la actualidad.

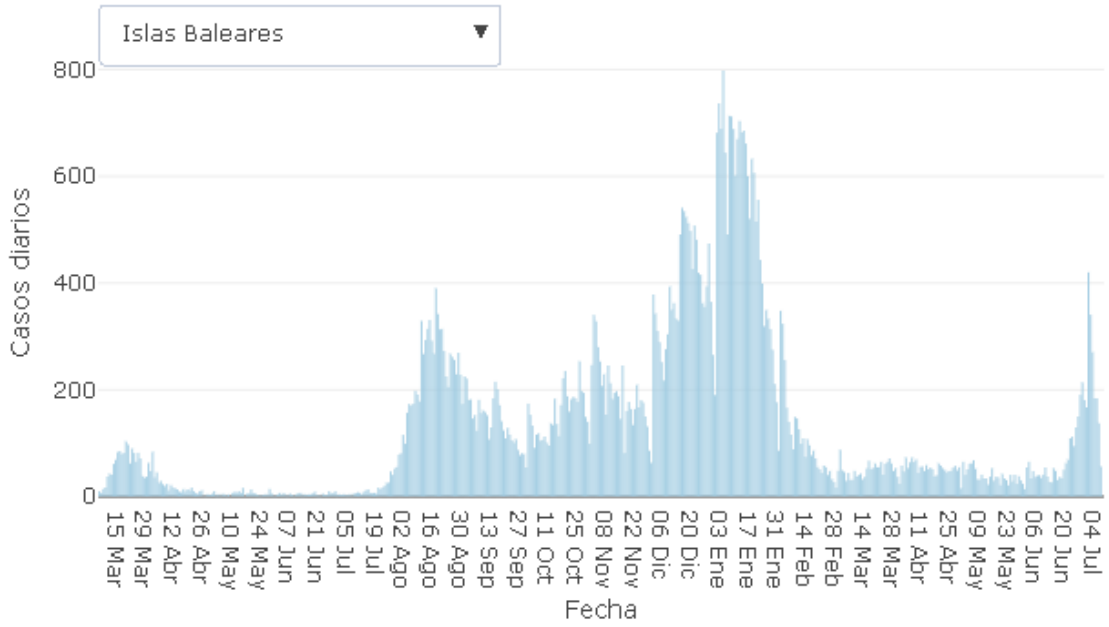

Figura 20-. Contagios diarios totales en las Islas Baleares desde que se tiene registros (15 de marzo) hasta la actualidad.

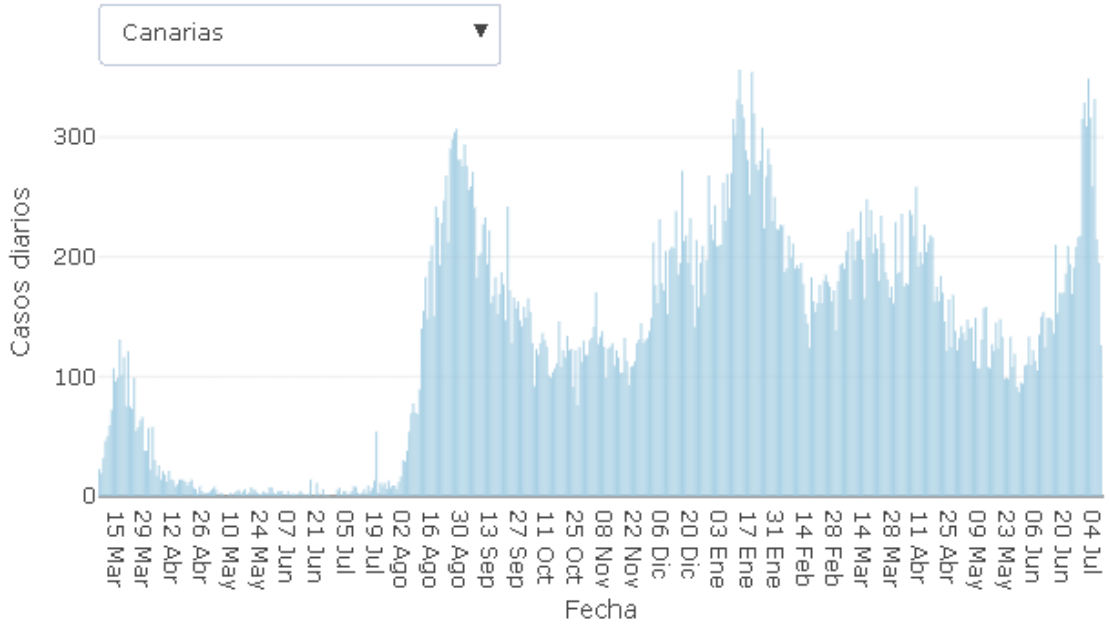

Figura 21-. Contagios diarios totales en las Islas Canarias desde que se tiene registros (15 de marzo) hasta la actualidad.

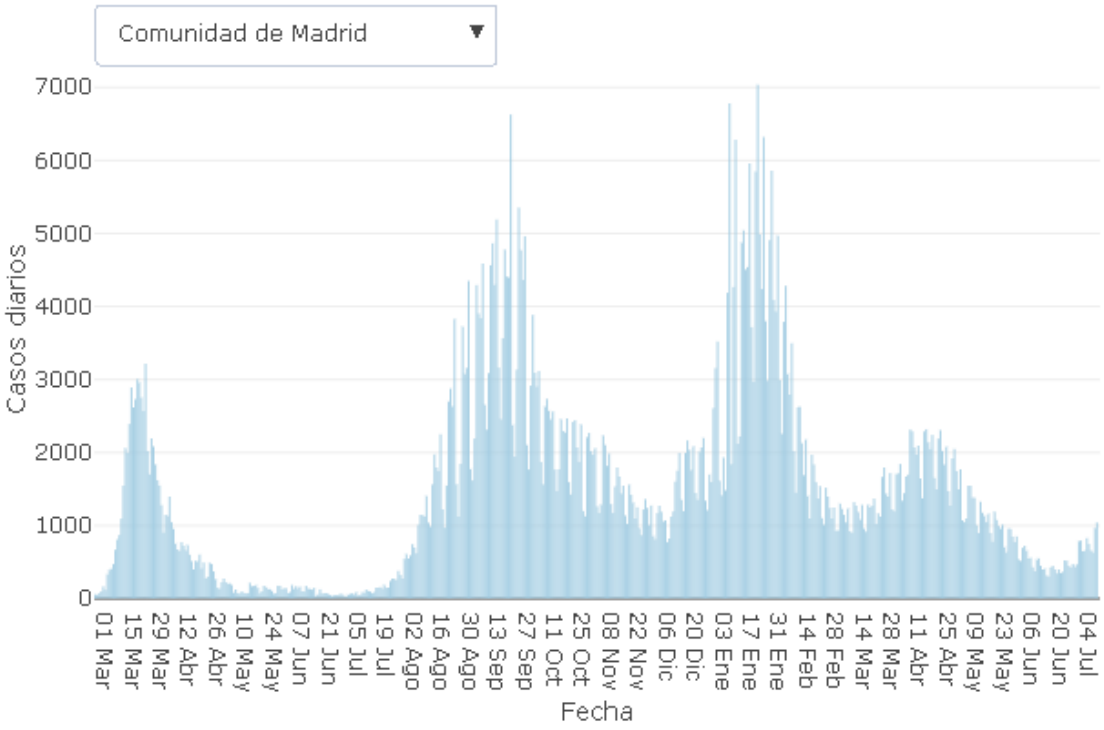

Figura 22-. Contagios diarios totales en la Comunidad de Madrid desde que se tiene registros (1 de marzo) hasta la actualidad.

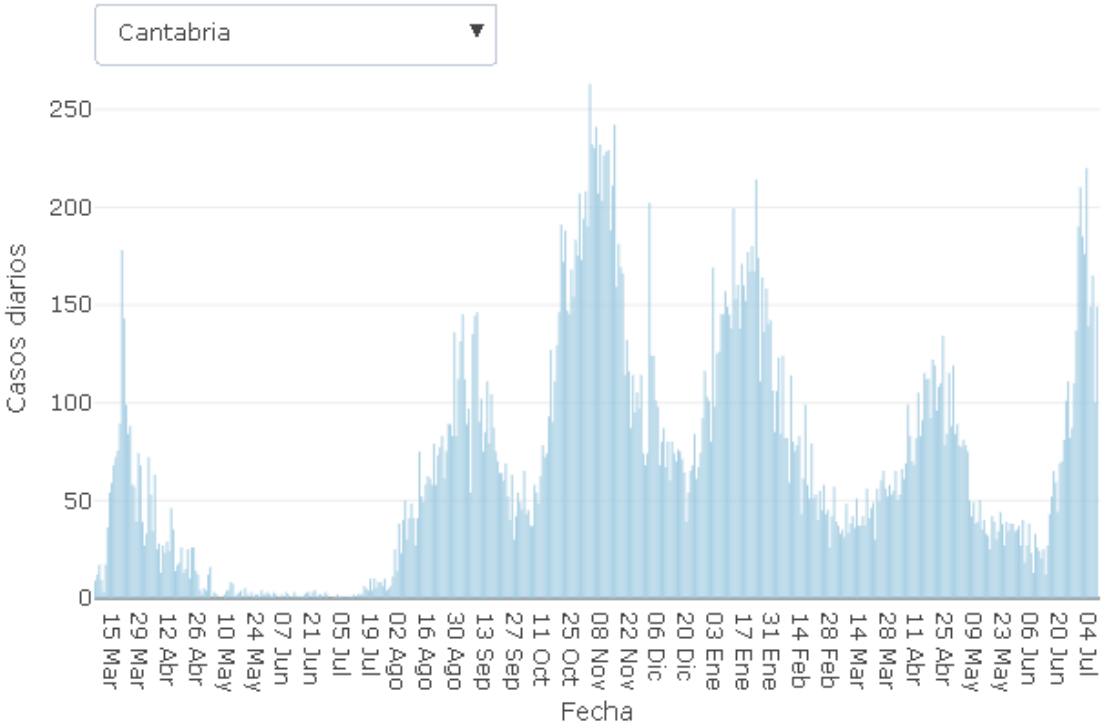

Figura 23-. Contagios diarios totales en Cantabria desde que se tiene registros (15 de marzo) hasta la actualidad.

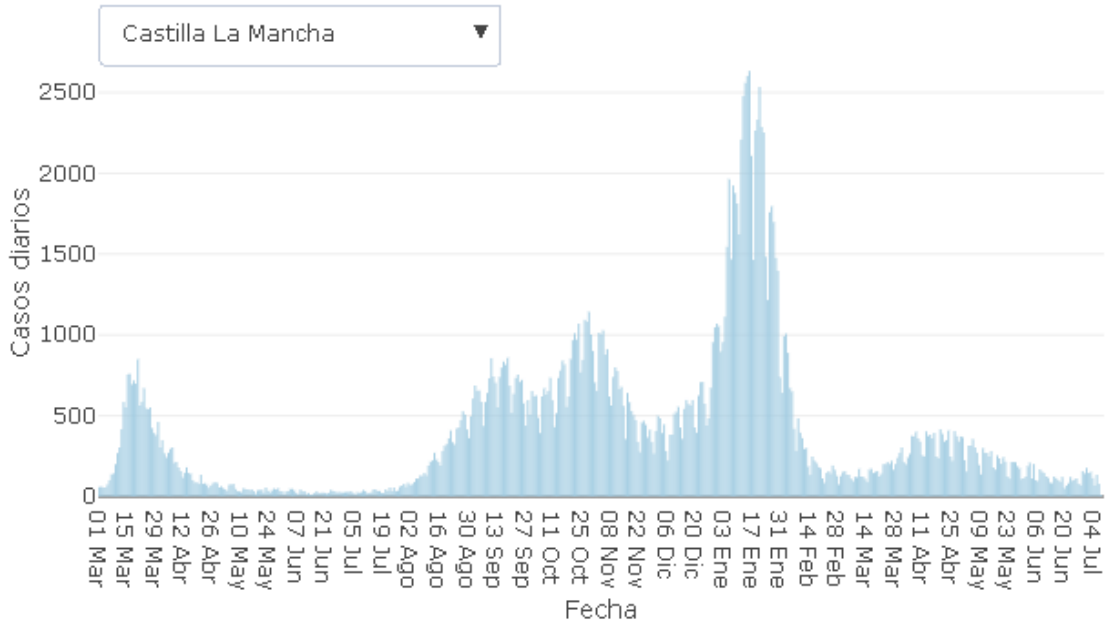

Figura 24-. Contagios diarios totales en Castilla La Mancha desde que se tiene registros (1 de marzo) hasta la actualidad.

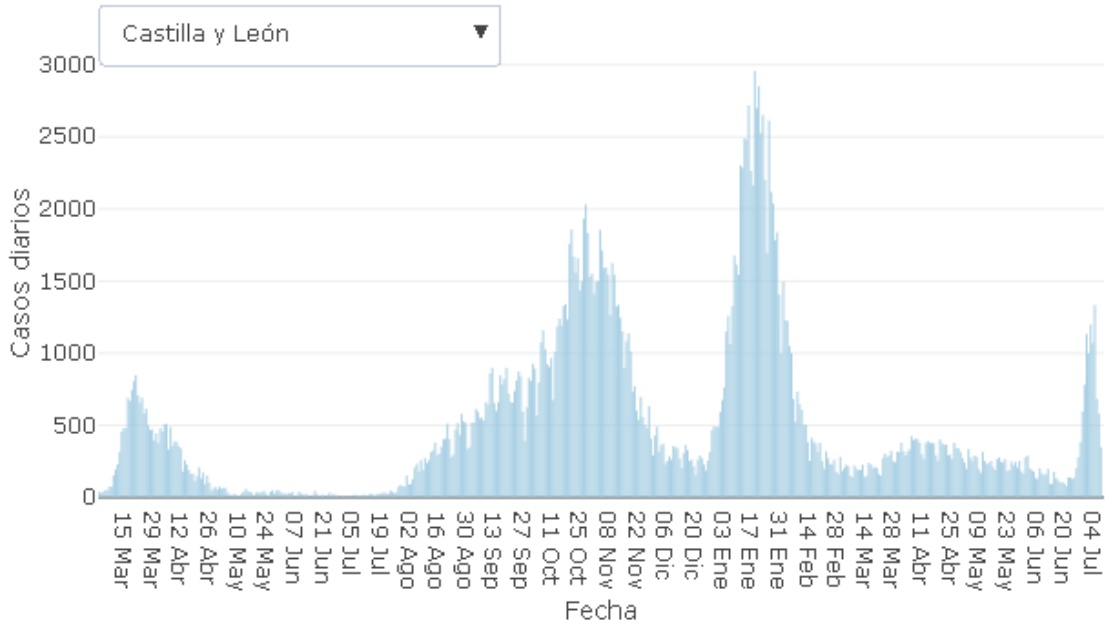

Figura 25-. Contagios diarios totales en Castilla y León desde que se tiene registros (15 de marzo) hasta la actualidad.

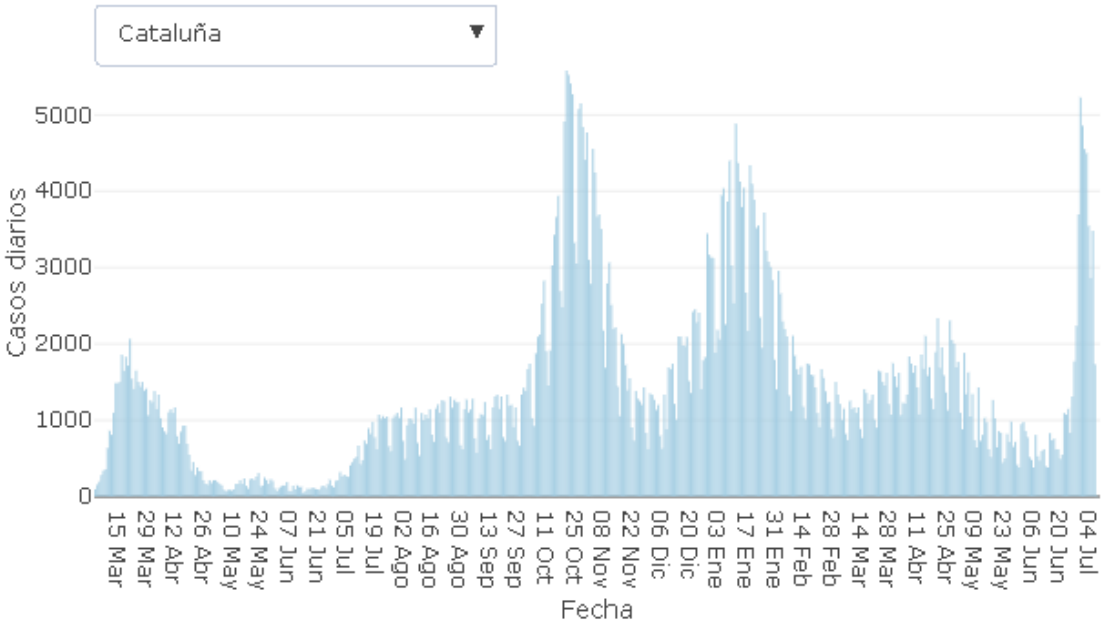

Figura 26-. Contagios diarios totales en Cataluña desde que se tiene registros (15 de marzo) hasta la actualidad.

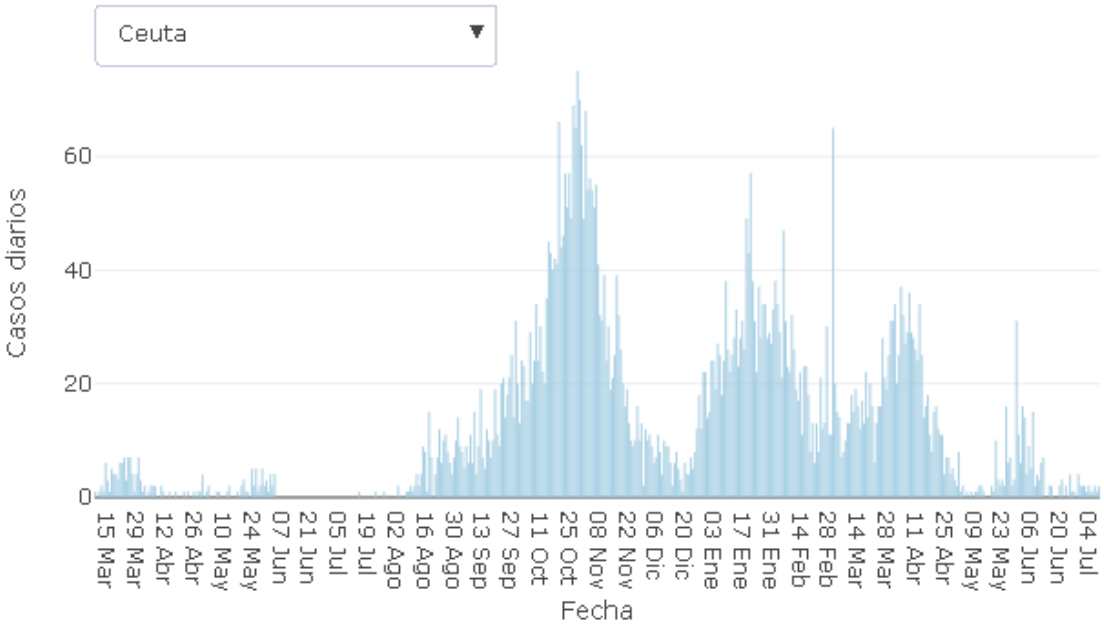

Figura 27-. Contagios diarios totales en Ceuta desde que se tiene registros (15 de marzo) hasta la actualidad.

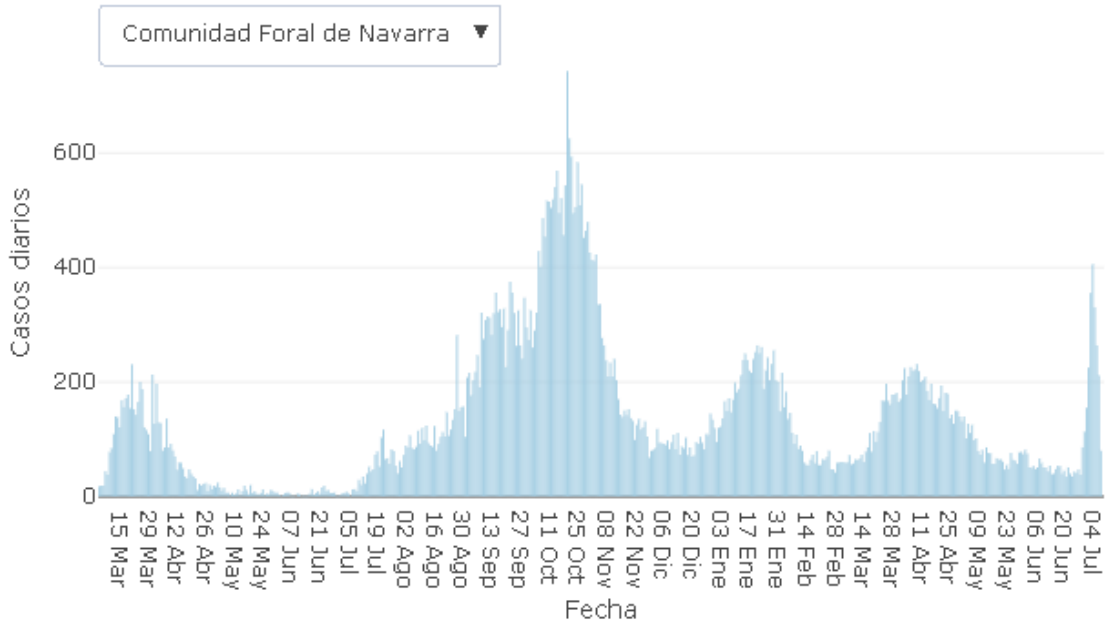

Figura 28-. Contagios diarios totales en la Comunidad Foral de Navarra desde que se tiene registros (15 de marzo) hasta la actualidad.

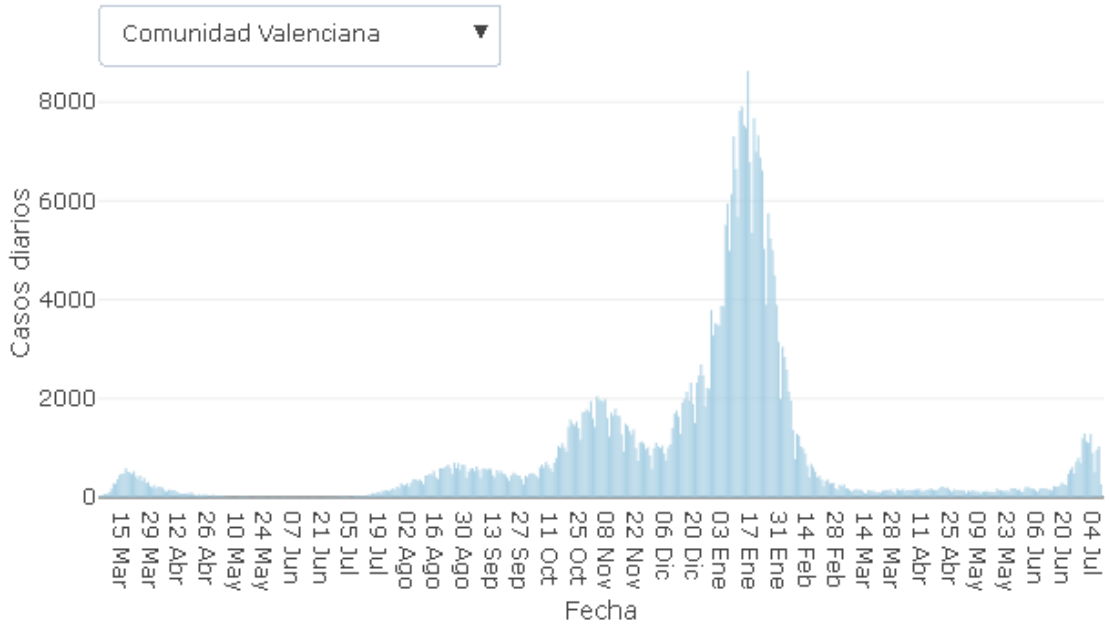

Figura 28-. Contagios diarios totales en la Comunidad Valenciana desde que se tiene registros (15 de marzo) hasta la actualidad.

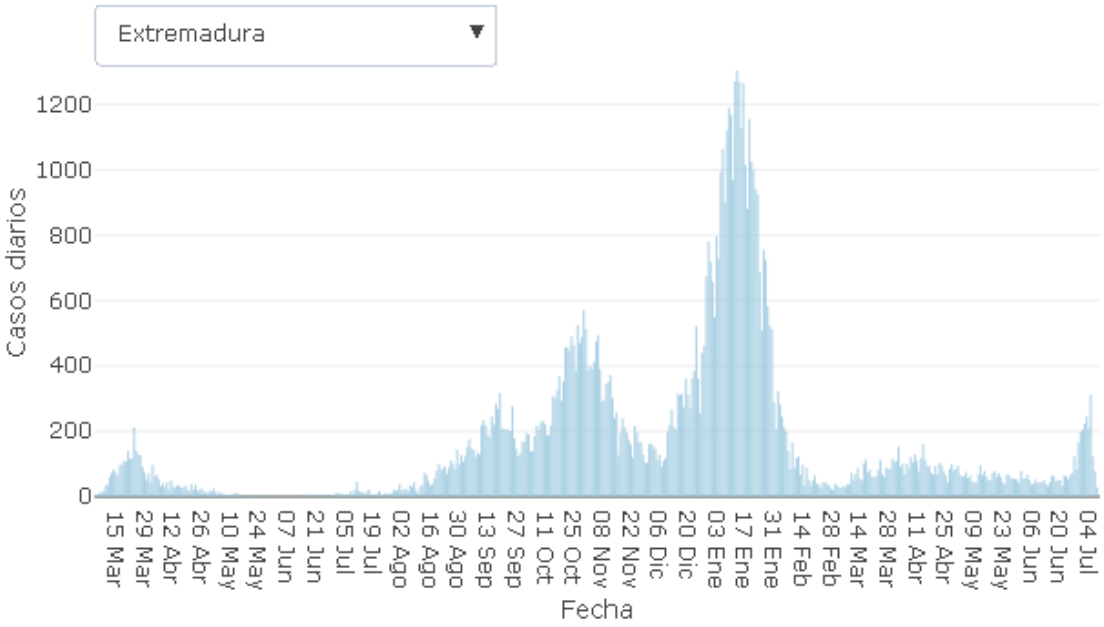

Figura 29-. Contagios diarios totales en Extremadura desde que se tiene registros (15 de marzo) hasta la actualidad.

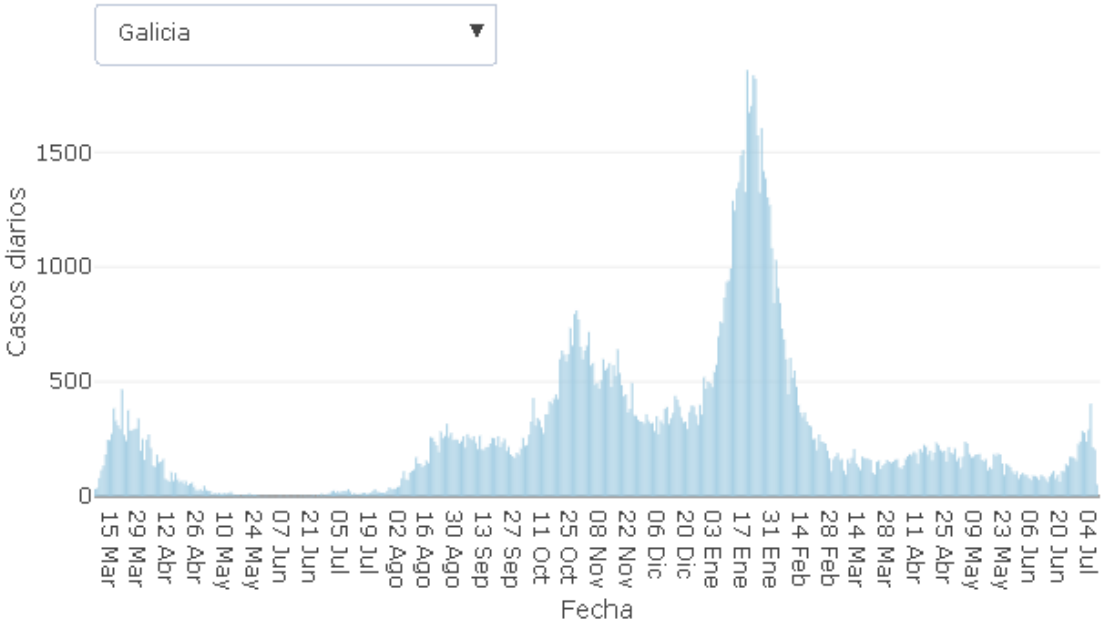

Figura 30-. Contagios diarios totales en Galicia desde que se tiene registros (15 de marzo) hasta la actualidad.

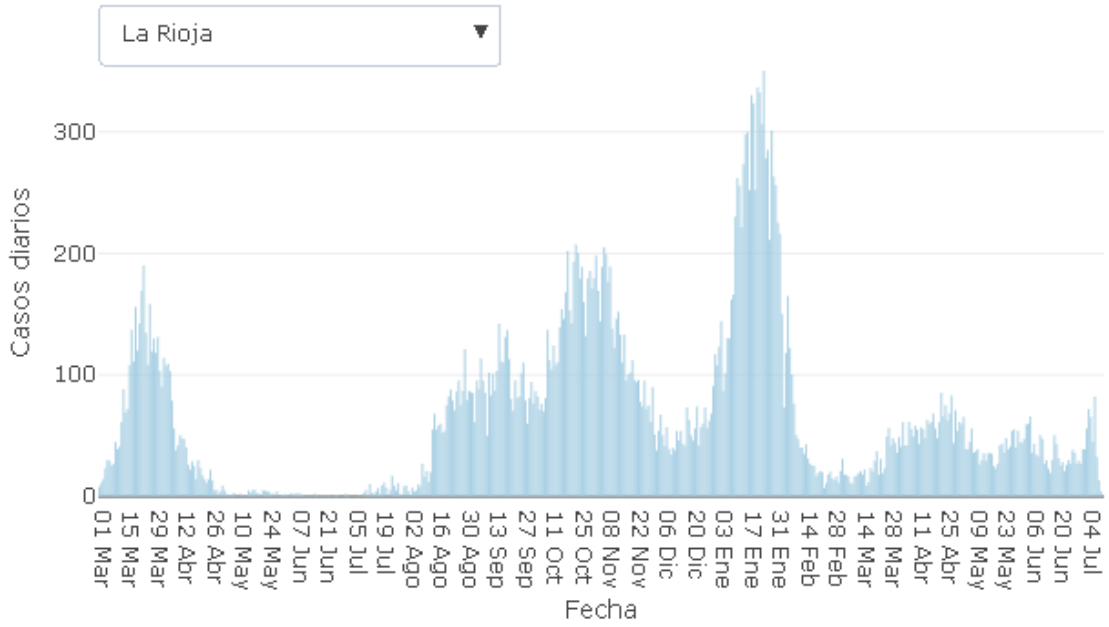

Figura 31-. Contagios diarios totales en La Rioja desde que se tiene registros (1 de marzo) hasta la actualidad.

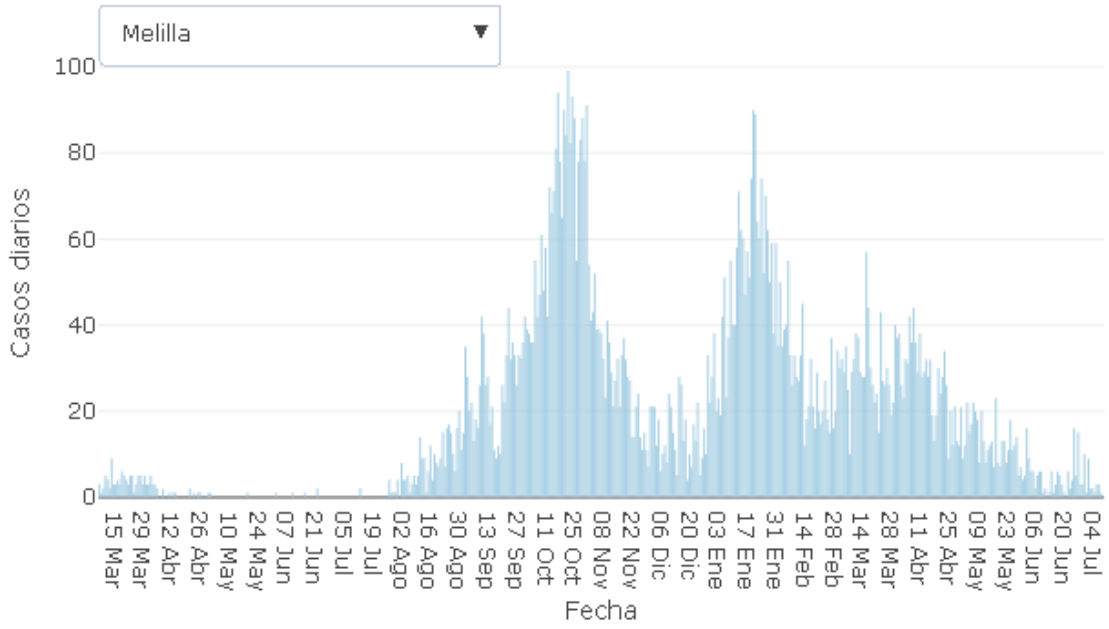

Figura 32-. Contagios diarios totales en Melilla desde que se tiene registros (15 de marzo) hasta la actualidad.

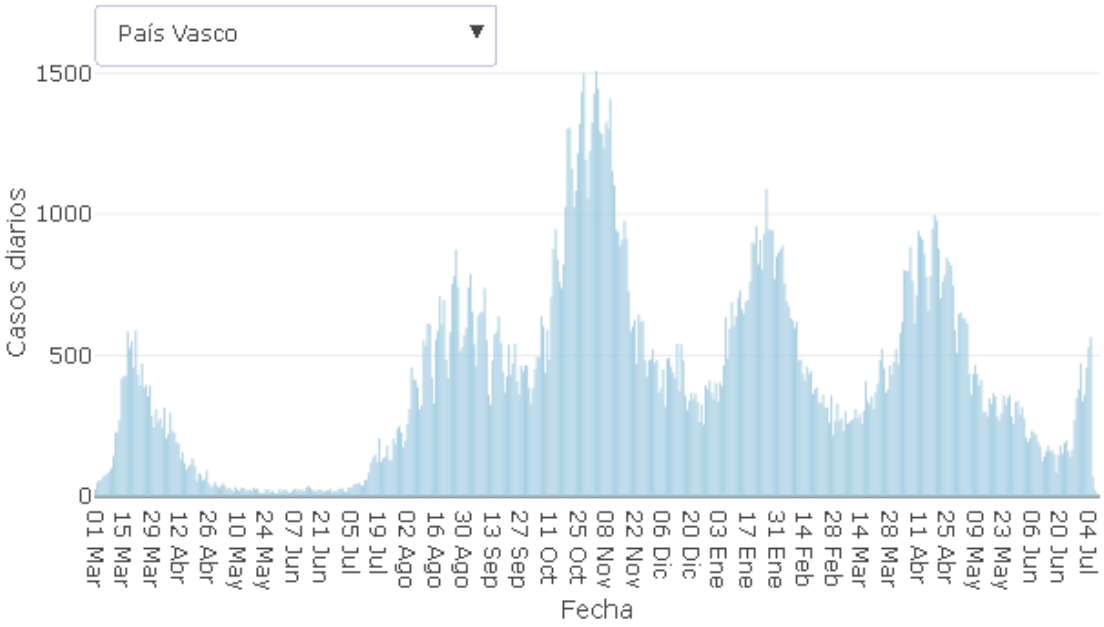

Figura 33-. Contagios diarios totales en el País Vasco desde que se tiene registros (15 de marzo) hasta la actualidad.

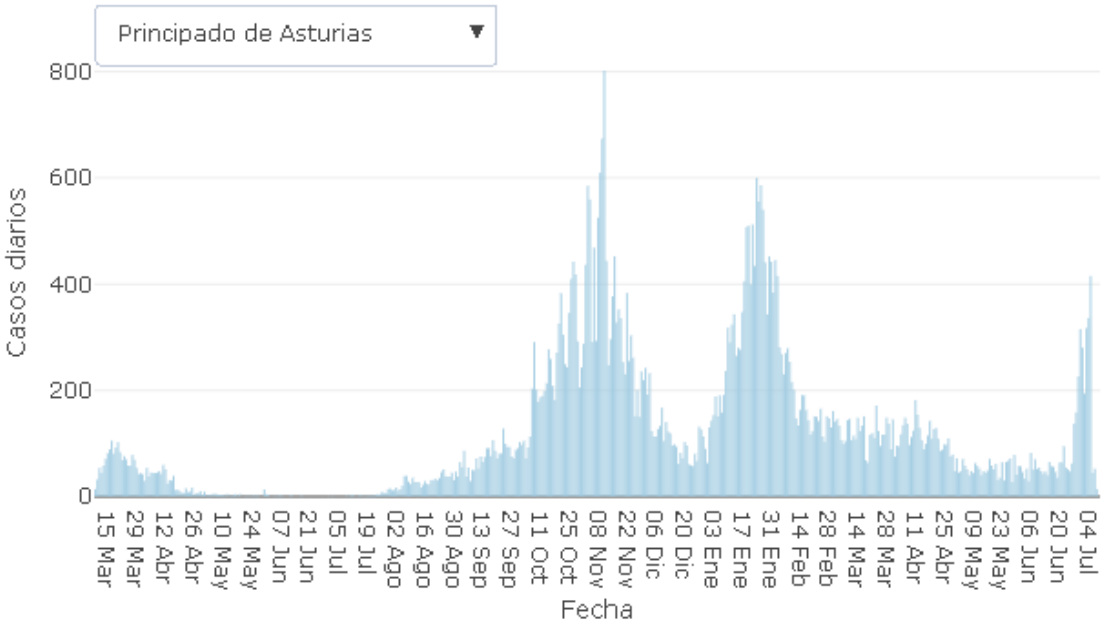

Figura 34-. Contagios diarios totales en el Principado de Asturias desde que se tiene registros (15 de marzo) hasta la actualidad.

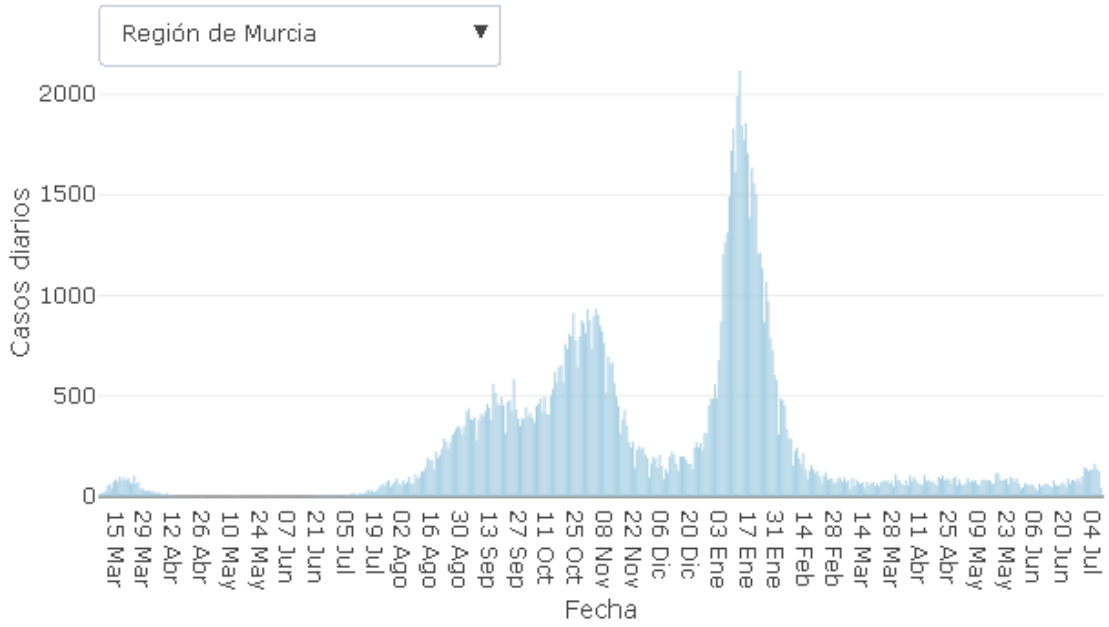

Figura 35-. Contagios diarios totales en la Región de Murcia desde que se tiene registros (15 de marzo) hasta la actualidad.

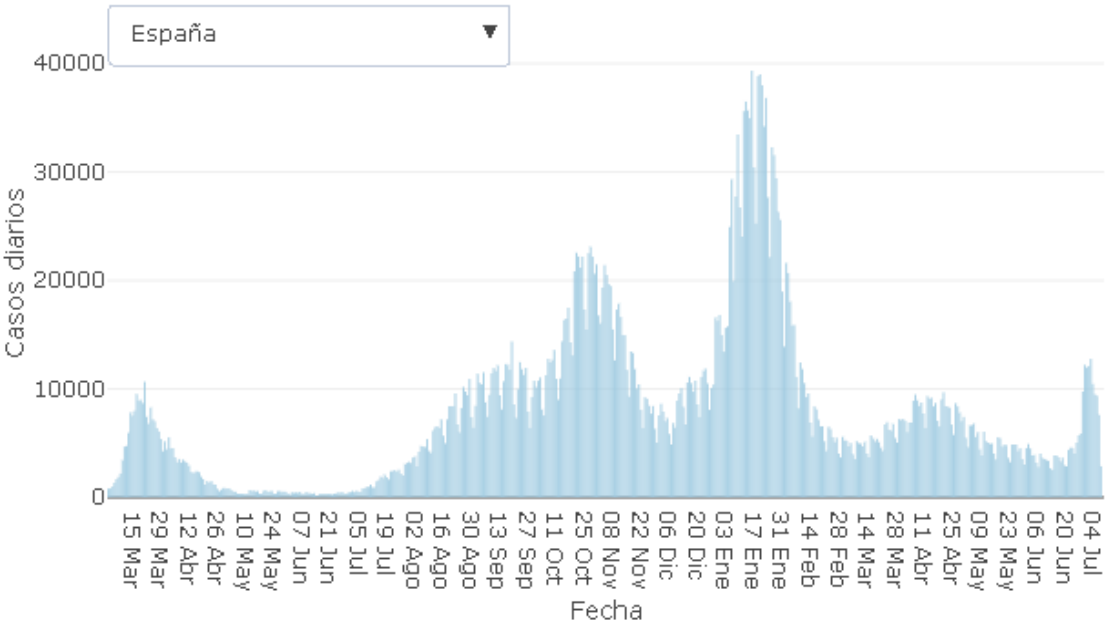

Figura 36-. Contagios diarios totales en España desde que se tiene registros (15 de marzo) hasta la actualidad.

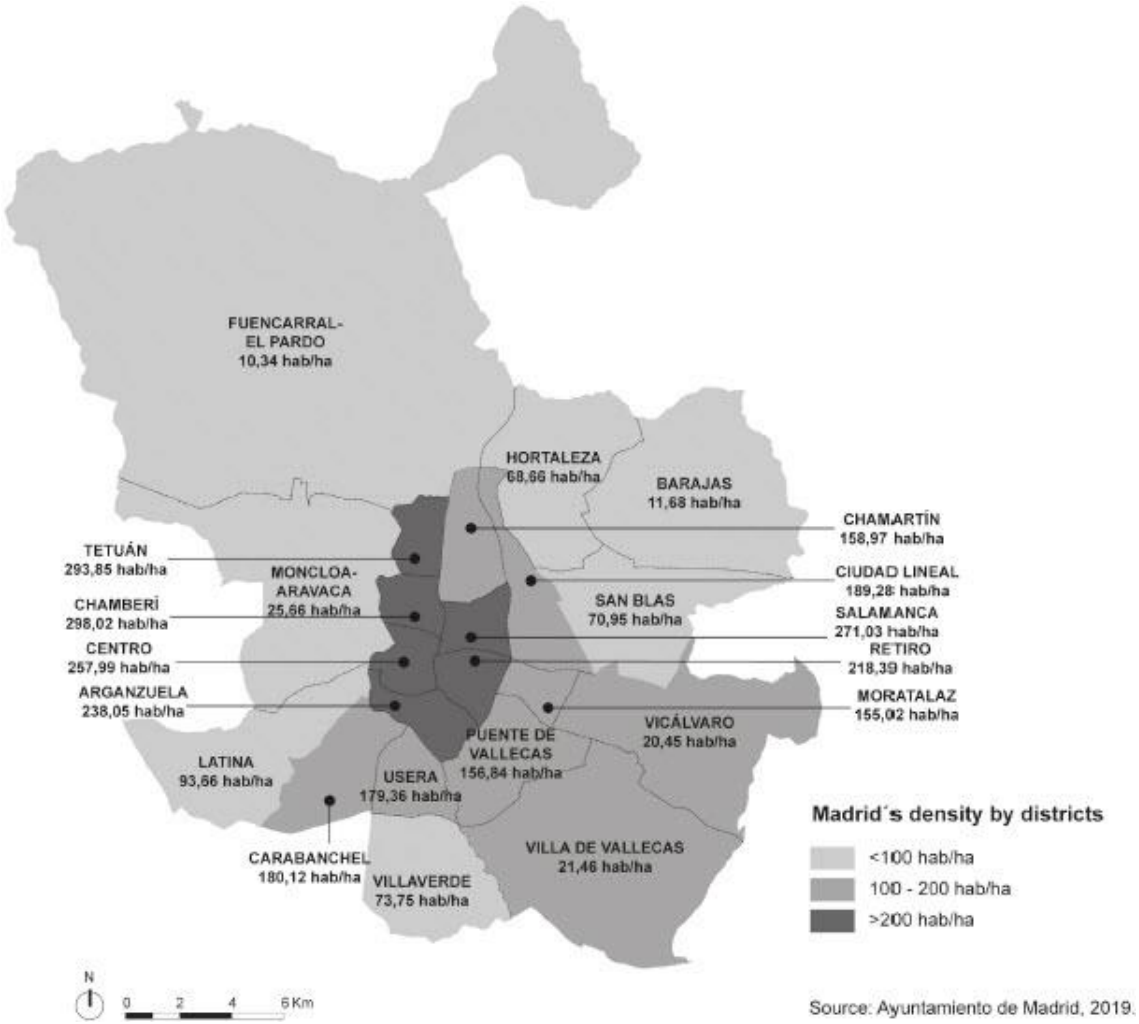

Figura 37-. Mapa representando la densidad poblacional de cada distrito de la ciudad de Madrid (Menéndez e Higueras, 2020).